Deviance Information Criterion for Latent Variable Models and Misspecified Models*

Yong Li
Renmin University of China

Jun Yu
Singapore Management University

Tao Zeng
Zhejiang University

October 31, 2019

Abstract

Deviance information criterion (DIC) has been widely used for Bayesian model comparison, especially after Markov chain Monte Carlo (MCMC) is used to estimate candidate models. This paper first studies the problem of using DIC to compare latent variable models when DIC is calculated from the conditional likelihood. In particular, it is shown that the conditional likelihood approach undermines theoretical underpinnings of DIC. A new version of DIC, namely DIC_L, is proposed to compare latent variable models. The large sample properties of DIC_L are studied. A frequentist justification of DIC_L is provided. Like AIC, DIC_L provides an asymptotically unbiased estimator to the expected Kullback-Leibler (KL) divergence between the DGP and a predictive distribution. Some popular algorithms, such as the EM, Kalman and particle filtering algorithms, are introduced to compute DIC_L for latent variable models. Moreover, this paper studies the problem of using DIC to compare misspecified models. A new version of DIC, namely DIC_M, is proposed and it can be regarded as a Bayesian version of TIC. A frequentist justification of DIC_M is provided under misspecification. DIC_L and DIC_M are illustrated using asset pricing models and stochastic volatility models.

JEL classification: C11, C12, G12
Keywords: AIC; TIC; DIC; Latent variable models; Misspecified models, Markov Chain Monte Carlo.

1 Introduction

Deviance information criterion (DIC) of Spiegelhalter et al. (2002) is a popular method for model selection in the Bayesian community. It has been used in a wide range of fields such

*We wish to thank two referees, an AE, Jianqing Fan, Eric Renault, Peter Phillips and David Spiegelhalter for their helpful comments. Yong Li, School of Economics and Institute of China's Reform & Development, Renmin University of China, Beijing, 100872, P.R. China. Jun Yu, School of Economics and Lee Kong Chian School of Business, Singapore Management University, 90 Stamford Rd, Singapore 178903. Email for Jun Yu: yujun@smu.edu.sg. URL: http://www.mysmu.edu/faculty/yujun/. Tao Zeng, School of Economics, Zhejiang University, Zhejiang, China 310058. Li gratefully acknowledges the financial support of the Chinese Natural Science Fund (No,71773130). Yu thanks the Singapore Ministry of Education for Academic Research Fund under grant number MOE2013-T3-1-009.

as biostatistics, ecology, etc. According to Spiegelhalter et al. (2014), Spiegelhalter et al. (2002) is the third most cited paper in international mathematical sciences between 1998 and 2008. Up to April 2019, it has received more than 5,800 citations on the Web of Knowledge and nearly 10,000 citations on Google Scholar. In economics and finance, DIC has received a lot of applications, for example, in stochastic frontier models (Galán et al., 2014), dynamic factors models (Bai and Wang, 2015), stochastic volatility models (Chan and Grant, 2016a, and Berg et al., 2004), and VAR models (Chan and Eisenstate, 2018).

The growth in popularity in DIC among applied researchers is understandable from a few aspects. First, DIC is a Bayesian version of the well-known Akaike Information Criterion (AIC) of Akaike (1973). Like AIC, DIC selects a model to minimize a plug-in predictive loss. This objective may appeal to applied researchers. Second, unlike AIC which is based on the log-likelihood function (or deviance) with the maximum likelihood (ML) estimate (MLE) of parameters being plugged in, DIC is based on the deviance with the posterior mean of parameters being plugged in. Li et al. (2017) gives the details about the loss functions associated with AIC and DIC. The detach of DIC from MLE is important when candidate models are difficult to estimate by ML. In this case, applied researchers may prefer Bayesian estimation methods over ML. In Bayesian statistics, the recent development of Markov chain Monte Carlo (MCMC) methods has been a key step in making it possible to estimate large hierarchical models. Large hierarchical models are typically difficult to estimate by ML, making ML-based model comparison criteria hard to implement. Third, DIC has a penalty term which can take account of prior information. This penalty term is different from that in AIC which only depends on the number of parameters in a candidate model.

Li et al. (2017) provided a frequentist justification to DIC by showing that DIC is an asymptotically unbiased estimator of the expected Kullback-Leibler (KL) divergence between the data generating process (DGP) and a predictive distribution with the posterior mean plugged in. The justification requires two critical assumptions. The first assumption is the validity of the Bernstein-von Mises theorem and the standard ML large sample theory (such as consistency and asymptotic normality). The second assumption is that all candidate models are asymptotically correctly specified. Both assumptions can be too strong in practice and hence, it is important to relax them.

This paper makes two contributions to the literature on DIC. First, we point out that the Bernstein-von Mises theorem and the standard ML large sample theory may not hold for the latent variables in latent variable models when DIC is calculated based on the conditional likelihood (i.e., the probability of observed data conditional on the original model parameter and the latent variables). We then propose a new version of DIC, namely DIC_L, in the context of latent variable models and provide a frequentist justification of DIC_L under some regularity conditions. We show that DIC_L is asymptotically equivalent to AIC when both are obtained the observed-data likelihood, that is, the likelihood with the latent variables being integrated

out. We also propose three methods to compute DIC_L in latent variable models.

Second, we propose a new version of DIC, namely DIC_M, for comparing misspecified models. We then provide a frequentist asymptotic justification of DIC_M and show that DIC_M is asymptotically equivalent to Takeuchi information criterion (TIC) of Takeuchi (1976).

The paper is organized as follows. Section 2 reviews DIC for model comparison. In Section 3, we review the widely-used DIC based on the conditional likelihood for comparing latent variable models and explain why the Bernstein-von Mises theorem may not hold for latent variables. We also introduce DIC_L based on the integrated likelihood for comparing latent variable models. Large sample properties of DIC_L are studied and several general algorithms are introduced to compute DIC_L in this section. Section 4 introduces DIC_M for misspecified models and obtains large sample relationships between DIC_M and TIC. Section 5 illustrates the methods using asset pricing models and stochastic volatility models. Section 6 concludes the paper. The Appendix collects the proof of theoretical results in the paper. An online supplement proves two statements in Remark 4.2.

2 DIC for Bayesian Model Comparison

Arguably the most important development in the Bayesian model comparison literature in recent years is DIC of Spiegelhalter et al. (2002). Compared with Bayes factors (BFs) which compare models through their "posterior probabilities" and try to search for the "true" model, DIC tries to find a better model for making "prediction" of replicate data.

DIC enjoys several desirable features. First, DIC is easy to calculate when the likelihood function has a closed-form expression and the posterior distribution is obtained by MCMC. Second, it applies to a wide range of statistical models. Third, unlike BFs, it is not subject to the Jeffreys-Lindley paradox and can be used when improper priors are used.

Consider a candidate parametric model, M, denoted by $p(\mathbf{y}|M,\theta)$ which is used to fit the data $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$, where θ is the parameter with P dimensions and $\theta \in \mathbf{\Theta} \subseteq R^P$. We will write $p(\mathbf{y}|M,\theta)$ as $p(\mathbf{y}|\theta)$ when there is no confusion. Letting $D(\theta) = -2\ln p(\mathbf{y}|\theta)$, DIC of Spiegelhalter et al. (2002) is given by

$$\text{DIC} = D\left(\bar{\theta}\right) + 2P_D, \tag{1}$$

where $\bar{\theta}$ is the posterior mean of θ, and P_D , known as "effective number of parameters", is given by:

$$P_D = -2\int \left[\ln p(\mathbf{y}|\theta) - \ln p(\mathbf{y}|\bar{\theta})\right] p(\theta|\mathbf{y})\mathrm{d}\theta. \tag{2}$$

Spiegelhalter et al.interprets $D\left(\bar{\theta}\right)$ as the Bayesian measure of model fit and P_D as the penalty term to measure model complexity.

DIC and AIC have some important differences. First, AIC is based on the MLE, while DIC is based on the posterior mean. Second, in AIC the penalty term depends on the number

of parameters, P, which is used to measure the model complexity. Hence, it is invariant to the prior. When the prior is informative, it imposes additional restrictions on the parameter space. In DIC the penalty term is determined by P_D whose value may depend on the prior. P_D may not be same as P in finite samples. As commented by Brooks (2002), an important contribution of DIC is to provide a way to measure the model complexity when an informative prior is used in a finite-sample setting.

Recently, under some mild regularity conditions, Li et al. (2017) provided a frequentist justification of DIC in the same manner as how AIC was justified. That is, both DIC and AIC try to find a model that asymptotically minimizes the expected KL divergence between the DGP and the corresponding predictive distribution. Other information criteria for comparing candidate models are possible. One example is Bayesian information criterion (BIC) of Schwarz (1978). More recently, Geweke and Amisano (2011) proposed a method that compares log predictive scores although, to the best of our knowledge, no general result is available on how to split samples when computing the log predictive scores. In Section 4.2, properties of AIC/DIC are compared with those of BFs/BIC. In this section, we first give a simple review of the justification of AIC/DIC.

Let $\mathbf{y}_{rep} = (y_{1,rep}, \cdots, y_{n,rep})$ be the independent replicate data of n observations generated by the same mechanism that gives rise to the observed data $\mathbf{y}$ and $g(\mathbf{y})$ is the DGP. The quantity that measures the quality of the candidate model in terms of its ability to make predictions of replicate data is given by the following KL divergence between $g(\mathbf{y}_{rep})$ and $p(\mathbf{y}_{rep}|\mathbf{y})$:

$$\begin{aligned} & KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right] = E_{\mathbf{y}_{rep}}\left[\ln \frac{g\left(\mathbf{y}_{rep}\right)}{p\left(\mathbf{y}_{rep}|\mathbf{y}\right)}\right] = \int \left[\ln \frac{g\left(\mathbf{y}_{rep}\right)}{p\left(\mathbf{y}_{rep}|\mathbf{y}\right)}\right] g\left(\mathbf{y}_{rep}\right) \mathrm{d}\mathbf{y}_{rep} \\ = & \int \ln g\left(\mathbf{y}_{rep}\right) g\left(\mathbf{y}_{rep}\right) \mathrm{d}\mathbf{y}_{rep} - \int \ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right) g\left(\mathbf{y}_{rep}\right) \mathrm{d}\mathbf{y}_{rep}, \end{aligned} \tag{3}$$

where $p\left(\mathbf{y}_{rep}|\mathbf{y}\right)$ denote a generic predictive distribution. Clearly the first term is the same across all candidate models which is denoted by C. Thus,

$$KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right] = C - \int \ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right) g\left(\mathbf{y}_{rep}\right) \mathrm{d}\mathbf{y}_{rep}.$$

Let AIC:$= -2\ln p\left(\mathbf{y}|\hat{\theta}(\mathbf{y})\right) + 2P$ where $\hat{\theta}(\mathbf{y})$ is the MLE of θ based on $\mathbf{y}$. If one chooses $p\left(\mathbf{y}_{rep}|\mathbf{y}\right)$ in (3) to be the plug-in distribution $p\left(\mathbf{y}_{rep}|\hat{\theta}(\mathbf{y})\right)$, then it is well-known that (see, for example, Burnham and Anderson (2002)), under some regularity conditions,

$$\begin{aligned} & E_{\mathbf{y}}\left\{2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\hat{\theta}(\mathbf{y})\right)\right]\right\} = 2C + E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\hat{\theta}(\mathbf{y})\right)\right] \\ = & 2C + E_{\mathbf{y}}\left(-2\ln p\left(\mathbf{y}|\hat{\theta}(\mathbf{y})\right) + 2P\right) + o(1) = 2C + E_{\mathbf{y}}\left(\mathrm{AIC}\right) + o(1), \end{aligned} \tag{4}$$

where the expectation $E_{\mathbf{y}}$ and $E_{\mathbf{y}_{rep}}$ are related to $g\left(\mathbf{y}\right)$ and $g\left(\mathbf{y}_{rep}\right)$, respectively. Hence, AIC is an asymptotically unbiased estimator of the expected KL divergence minus $2C$, that

is,

$$E_{\mathbf{y}}\left\{2\times KL\left[g\left(\mathbf{y}_{rep}\right),p\left(\mathbf{y}_{rep}|\hat{\theta}(\mathbf{y})\right)\right]\right\}-2C:=EKL_{ML}. \tag{5}$$

If one chooses $p\left(\mathbf{y}_{rep}|\mathbf{y}\right)$ in (3) to be the plug-in distribution $p\left(\mathbf{y}_{rep}|\bar{\theta}(\mathbf{y})\right)$, where $\bar{\theta}(\mathbf{y})$ is the posterior mean of θ based on $\mathbf{y}$, Li et al. (2017) showed that

$$\begin{aligned} & E_{\mathbf{y}}\left\{2\times KL\left[g\left(\mathbf{y}_{rep}\right),p\left(\mathbf{y}_{rep}|\bar{\theta}(\mathbf{y})\right)\right]\right\}=2C+E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\theta}(\mathbf{y})\right)\right] \\ = \quad & 2C+E_{\mathbf{y}}\left(-2\ln p\left(\mathbf{y}|\bar{\theta}(\mathbf{y})\right)+2P_D\right)+o(1)=2C+E_{\mathbf{y}}\left(\mathrm{DIC}\right)+o(1). \end{aligned} \tag{6}$$

DIC is an asymptotically unbiased estimator of the expected KL divergence minus $2C$, that is,

$$E_{\mathbf{y}}\left\{2\times KL\left[g\left(\mathbf{y}_{rep}\right),p\left(\mathbf{y}_{rep}|\bar{\theta}(\mathbf{y})\right)\right]\right\}-2C:=EKL_{B}. \tag{7}$$

The smaller AIC/DIC, the better predictive performance of the candidate model. When the prior information is dominated by likelihood asymptotically, Li et al. (2017) also showed that DIC and AIC are asymptotically equivalent, that is,

$$\mathrm{DIC}=\mathrm{AIC}+o_p(1), P_D=P+o_p(1).$$

This explains why DIC is regarded as a Bayesian version of AIC.

When deriving the asymptotic theory given in (6), Li et al. (2017) imposed a set of regularity conditions. Essentially these conditions ensure the following key asymptotic properties. First, the Bernstein-von Mises theorem holds. That is, the posterior distribution converges to a normal distribution with the MLE as its mean and the inverse of the second derivative of the negative log-likelihood function evaluated at the MLE as its covariance. In addition, the standard large sample theory for ML holds, including consistency, asymptotic normality with the covariance being the inverse of the second derivative of the negative log-likelihood function evaluated at the true parameter value. Second, all candidate models are correctly specified, at least asymptotically.

Unfortunately, the Bernstein-von Mises theorem and the standard large sample theory for ML may not hold for latent variables in many latent variable models. Moreover, the assumption that all candidate models are asymptotically correctly specified is too strong. In Section 4 we deal with the latent variables models and in Section 5 we relax the assumption of correct model specification.

3 DIC for Latent Variable Models

3.1 MCMC and data augmentation

A typical hierarchical model used in economics and finance involves latent variables. Latent variables have figured prominently in consumption decision, investment decision, labor force participation, conduct of monetary policy, indices of economic activity, inflation dynamics,

and other economic, business and financial activities and decisions. Not surprisingly, latent variable models have been widely used in financial econometrics, macroeconometrics and microeconometrics. For example, in financial econometrics it is often found that values of stocks, bonds, options, futures, and derivatives are often determined by a small number of factors. These factors, such as the level, the slope and the curvature in the term structure of interest rates, are latent. In macroeconomics, a well-known recent example of latent variable models is the dynamic factor model. Based on macroeconomic theory, the dynamic factor model attempts to explain aggregate economic phenomena by taking into account the fact that the economy is affected by some important factors. In microeconometrics, many discrete choice models and panel data models involve unobserved variables to capture observed heterogeneity across economic entities (Norets, 2009; Stern, 1997).

Let $\mathbf{y}$ be the observed data and $\mathbf{z} = (z_1, z_2, \cdots, z_n)'$ be the latent variables.[1] Let a latent variable model be indexed by a set of P parameters, $\theta \in \mathbf{\Theta} \subseteq R^P$. Let $p(\mathbf{y}|\theta)$ be the likelihood function of the observed data (denoted the observed-data likelihood), and $p(\mathbf{y}, \mathbf{z}|\theta)$ be the complete-data likelihood function. The relationship between the two functions is:

$$p(\mathbf{y}|\theta) = \int p(\mathbf{y}, \mathbf{z}|\theta) \mathrm{d}\mathbf{z}. \tag{8}$$

Typically the integral in (8) does not have a closed-form solution. Consequently, the ML method and hence, AIC are difficult to use as it requires calculations of $p(\mathbf{y}|\theta)$ for each value of θ during numerical optimizations.

If the Bayesian posterior analysis is conducted based on the observed-data likelihood, $p(\mathbf{y}|\theta)$, one would end up with the same problem as in ML since $p(\mathbf{y}|\theta)$ does not have a closed-form expression and, hence, the calculation of $\ln p(\mathbf{y}|\theta)$ for each MCMC draw is very time-consuming. An alternative way to conduct the Bayesian posterior analysis is based on $p(\mathbf{y}|\theta, \mathbf{z})$ (i.e. the conditional likelihood) which is often available in closed-form. In the conditional likelihood, we treat $\mathbf{z}$ in the same way as θ. In the Bayesian literature, this parameter expansion technique based on $p(\mathbf{y}|\theta, \mathbf{z})$ is known as data augmentation; see Tanner and Wong (1987) for further details. The closed-form expression of $p(\mathbf{y}|\theta, \mathbf{z})$ greatly facilitates MCMC sampling from the joint posterior distribution $p(\theta, \mathbf{z}|\mathbf{y})$. After a sufficiently long period for a burn-in phase, the simulated random samples can be regarded as random observations from the joint distribution. The statistical analysis can be established from these simulated posterior random observations. As a by-product to the Bayesian analysis, one also obtains MCMC samples for the latent variables $\mathbf{z}$. From the above discussion, it can be seen that data augmentation is the key technique for conducting the Bayesian posterior analysis of latent variable models, making MCMC a powerful alternative to ML as an estimation technique.

[1] Although we assume that the number of latent variables is the same as that of the observed data points, such an assumption may be relaxed. A more general assumption is that the number of latent variables grows proportionally with that of the observed data points. In this more general case, the theory discussed below continues to hold.

When the observed-data likelihood $p(\mathbf{y}|\theta)$ is not available in closed-form, DIC based on $p(\mathbf{y}|\theta)$ is very difficult to obtain, although the MCMC samples from $p(\theta, \mathbf{z}|\mathbf{y})$ are available. That explains why the widely-used DIC is obtained from the conditional likelihood $p(\mathbf{y}|\theta, \mathbf{z})$ but not from $p(\mathbf{y}|\theta)$ when there are latent variables in a candidate model. In fact, it is the default choice if one uses WinBUGS, a popular Bayesian software. As acknowledged in Spiegelhalter et al. (2014), this default way of calculating DIC from $p(\mathbf{y}|\theta, \mathbf{z})$ for latent variable models "is only to make the technique computationally feasible".

Unfortunately, when the DIC is calculated from $p(\mathbf{y}|\theta, \mathbf{z})$, the Bernstein-von Mises theorem and the standard ML large sample theory do not hold for latent variables. In fact, the posterior distribution of latent variables may not be normally distributed as the sample size goes to infinity. The posterior means of latent variables may not be close to the MLE even asymptotically. The MLE of latent variables may not be consistent. As a result, the asymptotic justification developed in Li et al. (2017) is no longer applicable.

The problem of calculating DIC from $p(\mathbf{y}|\theta, \mathbf{z})$ has been pointed out in the literature. For example, Millar (2009) documented strong evidence of poor performance of DIC in negative binomial and Poisson-lognormal models using simulated data. He found that DIC almost always prefers the Poisson-gamma model instead of the Poisson-lognormal model, even when data are simulated from a Poisson-lognormal model. Millar and McKechnie (2014) documented strong evidence of poor performance of DIC in state-space models using simulated data. Chan and Grant (2016a, 2016b) showed that, in the context of stochastic volatility models, DIC tends to favor overfitted models using simulated data.

3.2 DIC for latent variable models

As described in Section 3.1, in a latent variable model, there are three types of variables, the observed data $\mathbf{y}$, the latent variables $\mathbf{z}$, and the parameters θ. In the frequentist framework, the likelihood function, $p(\mathbf{y}|\theta) = \int p(\mathbf{y}, \mathbf{z}|\theta)\mathrm{d}\mathbf{z}$, is clearly defined. In this case, only θ, not $\mathbf{z}$, is treated as parameters. In the Bayesian framework, however, three likelihood functions may be used, $p(\mathbf{y}|\theta)$, $p(\mathbf{y}, \mathbf{z}|\theta)$, and $p(\mathbf{y}|\theta, \mathbf{z})$ which correspond to the observed-data likelihood, the complete-data likelihood, and the conditional likelihood. Using the terminology of Celeux et al. (2006), DIC based on $p(\mathbf{y}|\theta)$ and $p(\mathbf{y}|\theta, \mathbf{z})$ can be written, respectively, as

$$\begin{aligned}
\mathrm{DIC}_1 &= -2\ln p(\mathbf{y}|E_{\theta|\mathbf{y}}(\theta)) + 2\left\{-2E_{\theta|\mathbf{y}}\left[\ln p(\mathbf{y}|\theta)\right] + 2\ln p(\mathbf{y}|E_{\theta|\mathbf{y}}(\theta))\right\} := D\left(\bar{\theta}\right) + 2P_{D,1}, \\
\mathrm{DIC}_7 &= -2\ln p(\mathbf{y}|E_{\theta,\mathbf{z}|\mathbf{y}}(\theta,\mathbf{z})) + 2\left\{-2E_{\theta,\mathbf{z}|\mathbf{y}}\left[\ln p(\mathbf{y}|\theta,\mathbf{z})\right] + 2\ln p(\mathbf{y}|E_{\theta,\mathbf{z}|\mathbf{y}}(\theta,\mathbf{z}))\right\} := D\left(\bar{\theta},\bar{\mathbf{z}}\right) + 2P_{D,7}.
\end{aligned}$$

DIC_1 is monitored and reported in WinBUGS when there is no latent variable. To compute DIC_1, we approximate $E_{\theta|\mathbf{y}}\left[\ln p(\mathbf{y}|\theta)\right]$ by $\frac{1}{J}\sum_{j=1}^{J}\ln p\left(\mathbf{y}|\theta^{(j)}\right)$. This approximation error can be made arbitrarily small for a large J. When $p(\mathbf{y}|\theta)$ be available in closed-form, $\frac{1}{J}\sum_{j=1}^{J}\ln p\left(\mathbf{y}|\theta^{(j)}\right)$ is easy to compute once the MCMC samples $\left\{\theta^{(j)}\right\}_{j=1}^{J}$ are available even when J is very large. When there is no latent variable, $p(\mathbf{y}|\theta)$ is often available in closed-form.

Unfortunately, for many latent variable models, such as state-space models, $p(\mathbf{y}|\theta)$ is not available in closed-form. In this case, DIC_1 is difficult to compute because it needs to evaluate $p(\mathbf{y}|\theta)$ for J times. Given that J is usually large, computing $\frac{1}{J}\sum_{j=1}^{J}\ln p\left(\mathbf{y}|\theta^{(j)}\right)$ without an analytical expression for $\ln p(\mathbf{y}|\theta)$ is time-consuming, making $D\left(\bar{\theta}\right)$ and especially $P_{D,1}$ difficult to obtain. In DIC_7, the latent variables are regarded as parameters, and $\ln p(\mathbf{y}|\theta,\mathbf{z})$ often has an analytical expression. Hence, it is easy to compute $\frac{1}{J}\sum_{j=1}^{J}\ln p\left(\mathbf{y}|\mathbf{z}^{(j)},\theta^{(j)}\right)$ once the MCMC samples $\left\{\theta^{(j)},\mathbf{z}^{(j)}\right\}_{j=1}^{J}$ are available. Clearly $\frac{1}{J}\sum_{j=1}^{J}\ln p\left(\mathbf{y}|\mathbf{z}^{(j)},\theta^{(j)}\right)$ can arbitrarily well approximate $D\left(\bar{\theta},\bar{\mathbf{z}}\right)$ for large J. That is why, when there are latent variables, data augmentation is used to obtain Markov chains for both $\mathbf{z}$ and θ. Following the suggestion of Spiegelhalter et al. (2002), DIC_7 is monitored and reported in WinBUGS for latent variable models. Clearly the use of DIC_7 is for computational convenience.

However, from a theoretical viewpoint, DIC_7 has a few problems. First and foremost, with data augmentation, the dimension of the parameter space is much bigger, increasing from P to $n+P$. Since the dimension of the parameter space grows proportionally with the number of data points, the conditional likelihood $p(\mathbf{y}|\theta,\mathbf{z})$ is not regular, and it leads to the well-known incidental parameter problem in econometrics where information about these incidental parameters stops accumulating after a finite number of observations, often one, have been taken; see for example Neyman and Scott (1948) and Lancaster (2000). In this case the MLE is inconsistent. Similarly, the Bernstein-von Mises theorem becomes invalid; see Page 89-90 of Gelman et al. (2013). Therefore, DIC_7 lacks of frequentist justification. In fact, DIC_7 may not provide an asymptotically unbiased estimator of the KL divergence. For the same reason, if AIC is constructed based on $p(\mathbf{y}|\theta,\mathbf{z})$, then AIC would not provide an asymptotically unbiased estimator of the KL divergence.

To give an example where DIC_7 provide an asymptotically biased estimator of the KL divergence, let $y_i|\alpha_i,\sigma^2\sim N(\alpha_i,\sigma^2)$, $\alpha_i\sim N(0,1)$ for $i=1,...,n$. Clearly $y_i|\sigma^2\sim N(0,\sigma^2+1)$ and thus the MLE of σ^2 is $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}y_i^2-1$. It is straightforward to show $\hat{\sigma}^2$ is $\sqrt{n}$-consistent and asymptotically normally distributed. However, if $\{\alpha_i\}_{i=1}^{n}$ are treated as parameters, they are incidental in the sense of Neyman and Scott (1948). The MLE of α_i is $\hat{\alpha}_i=y_i\sim N(\alpha_i,\sigma^2)$ which is correctly centered at α_i but inconsistent as the variance of MLE does not go to zero as n grows. If $\sigma^2=1$ and is assumed to be known, then $P=n$ and the posterior distribution is $\alpha_i|y_i\sim N\left(0.5y_i,0.5\right)$. The posterior mean (which is also the posterior mode) is $\overline{\alpha}_i=0.5y_i$ which is not centered at the MLE. The posterior variance is 0.5 which does not go to zero as n grows. Clearly, both the standard ML large sample theory and the Bernstein-von Mises theorem fail to hold. These results are not surprising as only one observation (y_i) contains information about α_i.

Let $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_n)'$ and $\tilde{\alpha}(\mathbf{y})$ be an estimator of α. By evaluating (3) we have

$$KL\left[g\left(\mathbf{y}_{rep}\right),p\left(\mathbf{y}_{rep}|\tilde{\alpha}(\mathbf{y})\right)\right]=E_{\mathbf{y}_{rep}}\left[\ln\frac{g\left(\mathbf{y}_{rep}\right)}{p\left(\mathbf{y}_{rep}|\tilde{\alpha}\left(\mathbf{y}\right)\right)}\right]$$

$$
\begin{aligned}
&= C - \int \ln p\left(\mathbf{y}^{rep}|\tilde{\alpha}(\mathbf{y})\right) g(\mathbf{y}^{rep}) \mathrm{d}\mathbf{y}^{rep} \\
&= C + \left[\frac{n}{2} \ln(2\pi\sigma^2) + \frac{n\left(\sigma^2+1\right)}{2\sigma^2} + \sum_{i=1}^{n} \frac{\tilde{\alpha}_i^2(\mathbf{y})}{2\sigma^2} \right]. \qquad (9)
\end{aligned}
$$

When $\sigma^2 = 1$, by plugging the MLE of α_i (i.e., $\hat{\alpha}_i = y_i$) into (9), multiplying both sides by 2 and taking expectation with respect to $\mathbf{y}$, we have

$$
EKL_{ML} = n\ln(2\pi) + 2n + \sum_{i=1}^{n} E\left(y_i^2\right) = n\ln(2\pi) + 4n.
$$

However,

$$
E_{\mathbf{y}}(\mathrm{AIC}) = E_{\mathbf{y}}\left(-2\ln p\left(\mathbf{y}|\hat{\alpha}_1,\ldots,\hat{\alpha}_n\right)\right) + 2n = n\ln(2\pi) + 2n.
$$

Similarly, by plugging the posterior mean of α_i (i.e., $\overline{\alpha}_i = 0.5y_i$) into (9), multiplying both sides by 2 and taking expectation with respect to $\mathbf{y}$, we have

$$
EKL_B = n\ln(2\pi) + 2n + \sum_{i=1}^{n} \frac{E\left(y_i^2\right)}{4} = n\ln(2\pi) + 2.5n.
$$

However,

$$
\begin{aligned}
P_{D,7} &= -2\int \left[\ln p(\mathbf{y}|\alpha) - \ln p\left(\mathbf{y}|\overline{\alpha}_1,\ldots,\overline{\alpha}_n\right)\right] p(\alpha|\mathbf{y})\mathrm{d}\alpha \\
&= -2\int \left[\ln p(\mathbf{y}|\alpha)\right] p(\alpha|\mathbf{y})\mathrm{d}\alpha + 2\ln p\left(\mathbf{y}|\overline{\alpha}_1,\ldots,\overline{\alpha}_n\right) \\
&= \sum_{i=1}^{n} \int (y_i - \alpha_i)^2 p(\alpha_i|y_i)\mathrm{d}\alpha_i - \frac{\sum_{i=1}^{n} y_i^2}{2} \\
&= \sum_{i=1}^{n} \left[\frac{1}{2} + \frac{y_i^2}{4}\right] - \frac{\sum_{i=1}^{n} y_i^2}{2} = \frac{n}{2} - \frac{\sum_{i=1}^{n} y_i^2}{4},
\end{aligned}
$$

$$
\begin{aligned}
E_{\mathbf{y}}\left(\mathrm{DIC}_7\right) &= E_{\mathbf{y}}\left(-2\ln p\left(\mathbf{y}|\overline{\alpha}_1,\ldots,\overline{\alpha}_n\right) + 2P_D\right) \\
&= E_{\mathbf{y}}\left(n\ln(2\pi) + \frac{\sum_{i=1}^{n} y_i^2}{2} + 2P_D\right) = n\ln(2\pi) + n.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
EKL_{ML} &= E_{\mathbf{y}}(\mathrm{AIC}) + 2n, &(10)\\
EKL_B &= E_{\mathbf{y}}(\mathrm{DIC}_7) + 1.5n, &(11)\\
E_{\mathbf{y}}(P_{D,7}) &= 0 \neq n + o(1), &(12)\\
E_{\mathbf{y}}(\mathrm{AIC} - \mathrm{DIC}_7) &= n \neq o_p(1). &(13)
\end{aligned}
$$

According to (10) and (11), both AIC and DIC_7, if calculated from the conditional likelihood, provide asymptotically biased estimation to the corresponding expected KL divergence minus

$2C$. According to (12), on average the effective number of parameter ($P_{D,7}$) is zero. According to (13), on average AIC differs from DIC_7 by n. All these observations are at odds with the theory discussed earlier. The source of the problem lies in the presence of latent variables.

Second, sometimes a statistical model without latent variable can be represented by another model with latent variables. A leading example is the Student t distribution which can be rewritten as a normal-inverse-gamma distribution where the variance is assumed to follow an inverse-gamma distribution and hence, is treated as a latent variable. These two equivalent representations, even under the same priors, often lead to very different DIC values. The reason for this sharp discrepancy is that in the model without latent variables, DIC_1 is used while in the model with latent variables, DIC_7 is used. This problem arises in Section 8.2 of Spiegelhalter et al. (2002) and in Model 8 of Berg et al. (2004).

Third, due to data augmentation, the dimension of the parameter space becomes much larger and hence, DIC_7 is expected to be sensitive to transformations of latent variables. To illustrate this problem, we consider a simple transformation of latent variables in the well-known Clark model (Clark, 1973) which is given by,

$$\text{Model 1} : y_t \sim N(\mu, \exp(h_t)), h_t \sim N(0, \sigma^2), t = 1, \cdots, n. \tag{14}$$

An equivalent representation of the model is

$$\text{Model 2} : y_t \sim N(\mu, \sigma_t^2), \sigma_t^2 \sim LN(0, \sigma^2), t = 1, \cdots, n, \tag{15}$$

where LN denotes the log-normal distribution. In both models there are latent variables. In Model 2 the latent variable is the volatility σ_t^2 while the latent variable is the log-volatility $h_t = \ln \sigma_t^2$ in Model 1. Hence, following the usual practice in the literature, DIC_7 is the relevant version. Since the two models are identical, we expect the two models give the same DIC_7 value. To calculate DIC_7, we simulate 1000 observations from the model with $\mu = 0, \sigma^2 = 0.5$. Vague priors are selected for the two parameters, namely, $\mu \sim N(0, 100)$, $\sigma^{-2} \sim \Gamma(0.001, 0.001)$. We run Gibbs sampler to make 240,000 simulated draws from the posterior distributions. The first 40,000 are discarded as burn-in samples. The remaining observations with every 10th observation are collected as effective observations for statistical inference. With data augmentation, the latent variables, h_t and σ_t^2 are regarded as parameters, and we find that $P_{D,7} = 89.806$ and $\text{DIC}_7 = 2884.37$ for Model 1 but $P_{D,7} = 59.366$ and $\text{DIC}_7 = 2852.85$ for Model 2. These differences are very large. Given that we have the identical models and priors and use the same dataset, the vast differences suggest that DIC_7 and $P_{D,7}$ are very sensitive to transformations of latent variables.

To summarize the problems with DIC_7 in the context of latent variable models, while DIC_7 is easier to calculate and has been used widely in practice, it suffers from several theoretical problems. While DIC_1 has rigorously theoretical justification, it is very hard to compute from MCMC output since $p(\mathbf{y}|\theta)$ is not available in closed-form.

3.3 DIC_L for latent variable models

Based on the discussion above, there is a great need to introduce a new Bayesian model selection criterion which has a valid justification and applies to general latent variable models and feasible to compute. In this section, we propose a new version of DIC, DIC_L.

When $p(\mathbf{y}_{rep}|\mathbf{y})$ in (3) is chosen to be the plug-in distribution $p\left(\mathbf{y}_{rep}|\bar{\theta}(\mathbf{y})\right)$, where $\bar{\theta}(\mathbf{y})$ is the posterior mean of θ (we simply write $\bar{\theta}(\mathbf{y})$ as $\bar{\theta}$ when there is no confusion), DIC_L is defined as,[2]

$$\text{DIC}_L = D(\bar{\theta}) + 2P_L, \tag{16}$$

$$P_L = \mathbf{tr}\left\{\mathbf{I}(\bar{\theta})V(\bar{\theta})\right\}, \tag{17}$$

where $\mathbf{tr}$ denotes the trace of a matrix and

$$\mathbf{I}(\theta) = -\frac{\partial^2 \ln p(\mathbf{y}|\theta)}{\partial\theta\partial\theta'}, V(\bar{\theta}) = E\left[\left(\theta-\bar{\theta}\right)\left(\theta-\bar{\theta}\right)'|\mathbf{y}\right].$$

Clearly, the leading term in DIC_L is the same as that in DIC_1. However, the penalty term in DIC_1 is $2P_D$ while it is $2P_L$ in DIC_L.

To justify DIC_L, we will develop the large sample properties under some regularity conditions in the same manner as how DIC_1 was justified by Li et al. (2017). In particular, we will show that DIC_L can approximate AIC, and P_L can approximate P. Moreover, we will show that DIC_L provides asymptotically unbiased estimation to the KL divergence minus $2C$.

Let $\mathbf{y}^t := (y_0, y_1, \ldots, y_t)$ for any $0 \le t \le n$ and $l_t\left(\mathbf{y}^t, \theta\right) = \ln p(\mathbf{y}^t|\theta) - \ln p(\mathbf{y}^{t-1}|\theta)$ be the log-likelihood for the t^{th} observation for any $1 \le t \le n$. When there is no confusion, we suppress $l_t\left(\mathbf{y}^t, \theta\right)$ as $l_t(\theta)$ so that $\ln p(\mathbf{y}|\theta) = \sum_{t=1}^n l_t(\theta)$.[3] And define $l_t^{(j)}(\theta)$ to be the j^{th} derivative of $l_t(\theta)$ and $l_t^{(j)}(\theta) = l_t(\theta)$ when $j = 0$. The L_p-norm of a random matrix X is defined as $\|X\|_p = \left(\sum_i \sum_j E|X_{ij}|^p\right)^{1/p}$, and $\|X\|$ denotes the Euclidean norm of the appropriate dimension. We introduce the following functions

$$\mathbf{s}(\mathbf{y}^t, \theta) := \frac{\partial \ln p(\mathbf{y}^t|\theta)}{\partial\theta} = \sum_{i=1}^t l_i^{(1)}(\theta), \ \mathbf{H}(\mathbf{y}^t, \theta) := \frac{\partial^2 \ln p(\mathbf{y}^t|\theta)}{\partial\theta\partial\theta'} = \sum_{i=1}^t l_i^{(2)}(\theta),$$

$$\mathbf{s}_t(\theta) := l_t^{(1)}(\theta) = \mathbf{s}(\mathbf{y}^t, \theta) - \mathbf{s}(\mathbf{y}^{t-1}, \theta), \ \mathbf{H}_t(\theta) := l_t^{(2)}(\theta) = \mathbf{H}(\mathbf{y}^t, \theta) - \mathbf{H}(\mathbf{y}^{t-1}, \theta),$$

$$\mathbf{B}_n(\theta) := Var\left[\frac{1}{\sqrt{n}}\sum_{t=1}^n l_t^{(1)}(\theta)\right], \bar{\mathbf{H}}_n(\theta) := \frac{1}{n}\sum_{t=1}^n \mathbf{H}_t(\theta),$$

$$\bar{\mathbf{J}}_n(\theta) := \frac{1}{n}\sum_{t=1}^n \mathbf{s}_t(\theta)\mathbf{s}_t(\theta)', \mathbf{H}_n(\theta) := \int \bar{\mathbf{H}}_n(\theta) g(\mathbf{y})\, d\mathbf{y}, \ \mathbf{J}_n(\theta) = \int \bar{\mathbf{J}}_n(\theta) g(\mathbf{y})\, d\mathbf{y}.$$

[2] To estimate DIC, DIC_L and DIC_M, one needs to estimate several population quantities. To ensure the sample counterparts of population quantities from MCMC draws converge, proper conditions are needed. For example, a sufficient condition, originally due to Meyn and Tweedie (2012), is the Harris ergodicity. For the sake of space, throughout this paper we assume MCMC draws are well-behaved and Harris ergodic.

[3] In the definition of log-likelihood, we ignore the initial condition $\ln p(y_0)$. For weakly dependent data, the impact of the initial condition is asymptotically negligible.

In this paper, as in Li et al. (2017), we impose the following regularity conditions.

Assumption 1: $\mathbf{\Theta} \subset R^P$ is compact.

Assumption 2: $\{y_t\}_{t=1}^{\infty}$ satisfies the strong mixing condition with the mixing coefficient $\alpha(m) = O\left(m^{\frac{-2r}{r-2}-\varepsilon}\right)$ for some $\varepsilon > 0$ and $r > 2$.

Assumption 3: For all t, $l_t(\theta)$ satisfies the standard measurability and continuity condition, and the eight-times differentiability condition on $F_{-\infty}^t \times \mathbf{\Theta}$ where $F_{-\infty}^t = \sigma(y_t, y_{t-1}, \cdots)$.

Assumption 4: For $j = 0, 1, 2$, for any $\theta, \theta' \in \mathbf{\Theta}$, $\left\| l_t^{(j)}(\theta) - l_t^{(j)}(\theta') \right\| \leq c_t^j\left(\mathbf{y}^t\right) \|\theta - \theta'\|$ in probability, where $c_t^j\left(\mathbf{y}^t\right)$ is a positive random variable with $\sup_t \left\| c_t^j\left(\mathbf{y}^t\right) \right\|_1 < \infty$ and $\frac{1}{n}\sum_{t=1}^n \left(c_t^j\left(\mathbf{y}^t\right) - E\left(c_t^j\left(\mathbf{y}^t\right)\right)\right) \xrightarrow{p} 0$.

Assumption 5: For $j = 0, 1, \ldots, 8$, there exists a function $M_t(\mathbf{y}^t)$ such that for all $\theta \in \mathbf{\Theta}$, $l_t^{(j)}(\theta)$ exists, $\sup_{\theta \in \mathbf{\Theta}} \left\| l_t^{(j)}(\theta) \right\| \leqslant M_t(\mathbf{y}^t)$, and $\sup_t \left\| M_t(\mathbf{y}^t) \right\|_{r+\delta} \leq M < \infty$ for some $\delta > 0$, where r is the same as that in Assumption 2.

Assumption 6: $\left\{ l_t^{(j)}(\theta) \right\}$ is L_2-near epoch dependent with respect to $\{\mathbf{y}_t\}$ of size -1 for $0 \leqslant j \leqslant 1$ and $-\frac{1}{2}$ for $j = 2$ uniformly on $\mathbf{\Theta}$.

Assumption 7: Let θ_n^p be the pseudo-true value that minimizes the KL loss between the DGP and the candidate model

$$\theta_n^p = \arg\min_{\theta \in \mathbf{\Theta}} \frac{1}{n} \int \ln \frac{g(\mathbf{y})}{p(\mathbf{y}|\theta)} g(\mathbf{y}) \mathrm{d}\mathbf{y},$$

where $\{\theta_n^p\}$ is the sequence of minimizers interior to $\mathbf{\Theta}$ uniformly in n and $\lim_{n\to\infty} \theta_n^p \in \mathrm{Int}(\mathbf{\Theta})$. For all $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup \sup_{\Theta \setminus N\left(\theta_n^p, \varepsilon\right)} \frac{1}{n} \sum_{t=1}^n \left\{ E\left[l_t(\theta)\right] - E\left[l_t\left(\theta_n^p\right)\right] \right\} < 0, \tag{18}$$

where $N(\theta_n^p, \varepsilon)$ is the open ball of radius ε around θ_n^p.

Assumption 8: The sequence $\{\mathbf{H}_n(\theta_n^p)\}$ is negative definite and the sequence $\{\mathbf{B}_n(\theta_n^p)\}$ is positive definite, both uniformly in n.

Assumption 9: $\mathbf{H}_n(\theta_n^p) + \mathbf{B}_n(\theta_n^p) = o(1)$.

Assumption 10: The prior density $p(\theta)$ is eight-times continuously differentiable, $p(\theta_n^p) > 0$ and $\int \|\theta\|^2 p(\theta) \mathrm{d}\theta < \infty$.

Remark 3.1 *Assumption 1 is the compactness condition. Assumption 2 and Assumption 6 imply weak dependence in y_t and l_t. The first part of Assumption 3 is the continuity condition. Assumption 4 is the Lipschitz condition for l_t first introduced in Andrews (1987) to develop the uniform law of large numbers for dependent and heterogeneous stochastic processes. Assumption 5 contains the dominance condition for l_t. Assumption 7 is the identification condition used in Gallant and White (1998). These assumptions are well-known primitive conditions for developing the ML theory, namely consistency and asymptotic normality, for dependent and heterogeneous data; see, for example, Gallant and White (1988) and Wooldridge (1994).*

Remark 3.2 *A measurable function of a mixing process is mixing if the function only depend on finite number of lagged values of the mixing process (Gallant and White, 1988). In most latent variable models, however, the likelihood function and the score function depend on the distant past or future of the process. Assumption 6 is used to control the dependence of the function; see Gallant and White (1998), Davidson (1992, 1993), de Jong (1997).*

Remark 3.3 *The eight-times differentiability condition in Assumption 3 and the domination condition for up to the eighth derivative of l_t in Assumption 5 are important to develop a high order stochastic Laplace expansion. In particular, as shown in Kass et al. (1990), these two conditions, together with the well-known consistency condition for ML given by (19) below, are sufficient for developing the Laplace expansion. This consistency condition requires that, for any $\varepsilon > 0$, there exists $K_1(\varepsilon) > 0$ such that*

$$\lim_{n\to\infty} \Pr\left(\sup_{\boldsymbol{\Theta}\setminus N\left(\theta_n^p,\varepsilon\right)} \frac{1}{n}\sum_{t=1}^{n} \left[l_t(\theta) - l_t\left(\theta_n^p\right) \right] < -K_1(\varepsilon) \right) = 1. \tag{19}$$

Our Assumption 7 is clearly more primitive than the consistency condition (19). In the following lemma, we show that Assumptions 1-7, including the identification condition (18), are sufficient to ensure (19) as well as the concentration condition around the posterior mode given by Chen (1985) and the concentration condition around the MLE given by Kim (1994, 1998). Together with Assumption 10, the concentration condition suggests that the stochastic Laplace expansion can be applied to the posterior distribution and the asymptotic normality of posterior distribution can be established.

Remark 3.4 *Assumption 9 gives the exact requirement for a good model. It generalizes the definition of "information matrix equality"; see White (1996). It was used in Li et al. (2017) to show that AIC and DIC provide asymptotically unbiased estimation to the KL divergence minus $2C$. However, as we will show soon, Assumption 9 is not required to establish the asymptotic equivalence between DIC and AIC.*

Remark 3.5 *Assumption 10 ensures the second moment of the prior is finite. As argued in Geweke and Keane (2001), such a condition typically leads to a finite second moment of posterior. Moreover, it implies that the prior is negligible asymptotically.*

Lemma 3.1 *If Assumptions 1-7 hold true, then Equation (19) holds. Furthermore, if Assumptions 1-7 hold true, for any $\varepsilon > 0$, there exists $K_2(\varepsilon) > 0$ such that*

$$\lim_{n\to\infty} \Pr\left(\sup_{\boldsymbol{\Theta}\setminus N\left(\hat{\theta},\varepsilon\right)} \frac{1}{n}\left[\sum_{t=1}^{n} l_t(\theta) - \sum_{t=1}^{n} l_t\left(\theta_n^p\right) \right] < -K_2(\varepsilon) \right) = 1. \tag{20}$$

Lemma 3.2 below gives a high order approximation to the posterior mean and the posterior variance based on a high order Laplace expansion. To apply the Laplace expansion, we need to fix more notations. For convenience of exposition, we let $\overline{\mathbf{H}}_n^{(j)}(\theta) = \frac{1}{n}\sum_{t=1}^n l_t^{(j)}(\theta)$ for $j = 3, 4, 5$. Let $\pi(\theta) = \ln p(\theta)$, $p^{(j)}(\theta)$, $\pi^{(j)}(\theta)$ be the jth order derivatives of $p(\theta)$, $\pi(\theta)$ for $j = 1, 2$, and $\hat{p}$, $\hat{\pi}$, $\hat{p}^{(j)}$ and $\hat{\pi}^{(j)}$ be the values of functions $p(\theta)$, $\pi(\theta)$, $p^{(j)}(\theta)$ and $\pi^{(j)}(\theta)$ evaluated at $\hat{\theta}(\mathbf{y})$. When there is no confusion, we write $\hat{\theta}(\mathbf{y})$ as $\hat{\theta}$.

Lemma 3.2 *Let $Var(\theta|\mathbf{y}) = E\left[(\theta - \bar{\theta})(\theta - \bar{\theta})'|\mathbf{y}\right]$ be the posterior variance of θ. Under Assumptions 1-8 and 10, it can be shown that*

$$\bar{\theta} = \hat{\theta} + \frac{1}{n}B_1^1 + \frac{1}{n^2}\left(B_2^1 - B_3^1\right) + O_p\left(\frac{1}{n^3}\right),$$
$$vec\left[Var(\theta|\mathbf{y})\right] = -\frac{1}{n}vec\left(\bar{\mathbf{H}}_n^{-1}(\hat{\theta})\right) + \frac{1}{n^2}(F_1 + F_2) + O_p\left(\frac{1}{n^3}\right),$$

where B_1^1 is defined in (53), B_2^1 defined in (55), $B_3^1 = B_1^1 \times B_4^1$, B_4^1 defined in (62), F_1 defined in (76), F_2 defined in (77) with vec denoting the column-wise vectorization of a matrix.

Remark 3.6 *Under the different regularity conditions, the Bernstein-von Mises theorem states that the posterior distribution converges to a normal distribution with the MLE as its mean and the inverse of the second derivative of the negative log-likelihood function evaluated at the MLE as its variance. Based on the Bernstein-von Mises theorem, when the parameter is one-dimensional, Ghosh and Ramamoorthi (2003) developed the similar results with Lemma 3.2 for the iid case. In particular, Ghosh and Ramamoorthi (2003) showed that*

$$\bar{\theta} - \hat{\theta} = o_p(n^{-1/2}), Var(\theta|\mathbf{y}) + \frac{1}{n}\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right) = o_p(n^{-1}).$$

Our Lemma 3.2 extends the results of Ghosh and Ramamoorthi (2003) in three aspects: (1) to the weakly dependent case; (2) to the multi-dimensional case; (3) giving the exact order of the first and second moments of the difference between the posterior distribution and the asymptotic normal distribution. From Lemma 3.2, we have

$$\bar{\theta} - \hat{\theta} = O_p(n^{-1}), Var(\theta|\mathbf{y}) + \frac{1}{n}\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right) = O_p(n^{-2}).$$

Based on this lemma, we can obtain the exact order of the difference between DIC_L and AIC as follows.

Theorem 3.1 *Under Assumptions 1-8 and 10, we have*

$$P_L = P + \frac{1}{n}C_1 + \frac{1}{n}C_2 + O_p\left(\frac{1}{n^2}\right),$$

$$DIC_L = AIC + \frac{1}{n}D_1 + \frac{1}{n}D_2 + O_p\left(\frac{1}{n^2}\right),$$

where

$$C_1 = \frac{1}{2}C_{11} - \frac{1}{2}C_{12}, \quad C_2 = -C_{22},$$

$$D_1 = C_{11} + \frac{5}{4}C_{12}, \quad D_2 = C_{21} - 2C_{22} - C_{23},$$

$$C_{11} = \mathbf{tr}\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right) \otimes vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\hat{\theta}\right)\right],$$

$$C_{12} = vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right),$$

$$C_{21} = \pi^{(1)}\left(\hat{\theta}\right)'\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right),$$

$$C_{22} = \mathbf{tr}\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\pi^{(2)}\left(\hat{\theta}\right)\right], \quad C_{23} = \pi^{(1)}\left(\hat{\theta}\right)'\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\pi^{(1)}\left(\hat{\theta}\right).$$

Corollary 3.2 *Under Assumptions 1-10, we have*

$$\begin{aligned} & E_{\mathbf{y}}\left\{2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\bar{\theta}\right)\right]\right\} = 2C + E_{\mathbf{y}}\left[-2\ln p\left(\mathbf{y}|\bar{\theta}\right) + 2P_L\right] + o(1) \\ = \;& 2C + E_{\mathbf{y}}\left(DIC_L\right) + o(1). \end{aligned}$$

Remark 3.7 *In Equation (15) on Page 590, Spiegelhalter et al. (2002) obtained the expression for P_L and claimed that P_L approximates P_D in DIC_1 and P in AIC. Unfortunately, to the best of our knowledge, P_L has never been implemented in practice, and WinBUGS does not report P_L. Moreover, the conditions under which $P_L \approx P_D \approx P$ holds true were not specified in Spiegelhalter et al. (2002). The order of the approximation error was unknown. According to Theorem 3.1, the difference between P and P_L and that between AIC and DIC_L are both $O_p(n^{-1})$. Furthermore, combined with Lemma 3.3 in Li et al. (2017), we can show that the approximation error between P_D and P_L and that between DIC_1 and DIC_L are both $O_p(n^{-1})$.*

Remark 3.8 *Without Assumption 9, Theorem 3.1 clearly shows that the difference between AIC and DIC_L is $O_p(n^{-1})$. For this reason, both DIC_L and DIC_1 can be regarded as the Bayesian version of AIC. When the prior is informative and the sample size is finite, DIC_L may give a different value from AIC. Like DIC_1, an important feature of DIC_L is that it provides an approach to measure the model complexity when the informative prior is available. According to Theorem 3.1, an alternative version of DIC, with or without latent variable, is $D(\bar{\theta}) + 2P$. In this case, the penalty term does not take into account of the prior information.*

Remark 3.9 *Corollary 3.2 is the direct result of Theorem 3.1 and Theorem 3.1 of Li et al. (2017). Since the frequentist justification of DIC and AIC needs Assumption 9, it is also needed to justify DIC_L as in Corollary 3.2. As DIC_1, DIC_L is an asymptotically unbiased estimator of the expected KL divergence minus $2C$. Hence, DIC_L selects a model that minimizes the expected KL divergence between the DGP and the plug-in predictive distribution. The smaller the value of DIC_L, the better the predictive performance of the candidate latent variable model.*

Remark 3.10 *From the discussion above, DIC_1 and DIC_L share the same asymptotic properties. However, as explained before, there is an important difference between DIC_1 and DIC_L, that is, the penalty term takes a different expression. It is this difference that makes DIC_L easier to compute from MCMC output. To compute $P_{D,1}$ in DIC_1, one has to evaluate $\frac{1}{J}\sum_{j=1}^{J} \ln p\left(\mathbf{y}|\theta^{(j)}\right)$ and hence calculate $p\left(\mathbf{y}|\theta^{(j)}\right)$ for J times. For latent variable models, since $p\left(\mathbf{y}|\theta^{(j)}\right)$ is not available in closed-form, the computational cost is high. However, to compute P_L in DIC_L, one needs to evaluate the second derivative of observed-data likelihood only once, which is computationally much less expensive. In Section 4.3, we will introduce some efficient algorithms to evaluate $D\left(\overline{\theta}\right)$ and $\mathbf{I}\left(\overline{\theta}\right)$.*

Remark 3.11 *In the context of latent variable models, while DIC_7 is trivial to calculate but cannot be justified, DIC_1 is justified but hard to compute. DIC_L solves this dilemma because it is justified and inexpensive to compute. The corresponding deviance is based on the observed-data likelihood function and the latent variables are not treated as parameters. It is important to point out that DIC_L is computed from MCMC output. While DIC_L does not treat latent variables as parameters, MCMC output may be obtained based on the data augmentation technique without affecting the asymptotic justification of DIC_L. Returning to the Clark model, with the same setting as before, we get $P_L = 1.75$ for Model 1 and $P_L = 1.80$ for Model 2. There is no significant difference between them. Moreover, these two values are close to 2, that is the actual number of parameters in the model. This result is what we expect given that the vague priors are used. The small difference between P_L and P arises due to the simulation error and the priors.*

3.4 Computing DIC_L for latent variable models

To calculate DIC_L, one needs to calculate $p(\mathbf{y}|\theta)$ and its derivatives with respect to θ (but there is no need to optimize $p(\mathbf{y}|\theta)$). Since there is no analytical expression for $p(\mathbf{y}|\theta)$ for many latent variable models, in this section, we show how to use the EM algorithm, the Kalman filter, and the particle filters to calculate $p(\mathbf{y}|\theta)$ and its derivatives with respect to θ.

3.4.1 Computing DIC_L by the EM algorithm

In this subsection we show how the EM algorithm may be used to evaluate $p(\mathbf{y}|\bar{\theta})$, the second derivative of the observed-data likelihood function, and hence DIC_L for the latent variable models. The EM algorithm is a powerful tool to deal with latent variable models. Instead of maximizing the observed-data likelihood function, the EM algorithm maximizes the so-called $\mathcal{Q}$ function given by

$$\mathcal{Q}(\theta|\theta^{(r)}) = E_{\theta^{(r)}}\{\mathcal{L}_c(\mathbf{y},\mathbf{z}|\theta)|\mathbf{y},\theta^{(r)}\}, \tag{21}$$

where $\mathcal{L}_c(\mathbf{y},\mathbf{z}|\theta) := \ln p(\mathbf{y},\mathbf{z}|\theta)$ is the complete-data likelihood function. The $\mathcal{Q}$ function is the conditional expectation of $\mathcal{L}_c(\mathbf{y},\mathbf{z}|\theta)$ with respect to the conditional distribution $p(\mathbf{z}|\mathbf{y},\theta^{(r)})$

where $\theta^{(r)}$ is a current fit of the parameter. The EM algorithm consists of two steps: the *expectation* (E) step and the *maximization* (M) step. The E-step evaluates $\mathcal{Q}(\theta|\theta^{(r)})$. The M-step determines a $\theta^{(r)}$ that maximizes $\mathcal{Q}(\theta|\theta^{(r)})$. Under some mild regularity conditions, for large enough r, $\{\theta^{(r)}\}$ obtained from the EM algorithm is the MLE, $\hat{\theta}$. For more details about the EM algorithm, see Dempster et al. (1977).

Although the EM algorithm is a good approach to dealing with latent variable models, the numerical optimization in the M-step is often unstable. Not surprisingly, the EM algorithm has been less popular to estimate latent variables models compared with MCMC techniques. However, we will show that, without numerical optimizations in the M-step, the theoretical properties of the EM algorithm facilitate computation of DIC_L for latent variable models.

It is noted that for any θ and θ^*in Θ, let $\mathcal{H}(\theta|\theta^*) = \int \ln p(\mathbf{z}|\mathbf{y},\theta)p(\mathbf{z}|\mathbf{y},\theta^*)\mathrm{d}\mathbf{z}$, the so-called $\mathcal{H}$ function in the EM algorithm. It was shown in that

$$\ln p(\mathbf{y}|\theta) = \mathcal{Q}\left(\theta|\theta^*\right) - \mathcal{H}\left(\theta|\theta^*\right).$$

Hence, $\ln p(\mathbf{y}|\bar{\theta})$ may be obtained as

$$\ln p(\mathbf{y}|\bar{\theta}) = \mathcal{Q}(\bar{\theta}|\bar{\theta}) - \mathcal{H}(\bar{\theta}|\bar{\theta}). \tag{22}$$

It can be seen that even when $\mathcal{Q}(\bar{\theta}|\bar{\theta})$ is not available in closed form, it is easy to evaluate from MCMC output because

$$\mathcal{Q}(\bar{\theta}|\bar{\theta}) = \int \ln p(\mathbf{y},\mathbf{z}|\bar{\theta})p(\mathbf{z}|\mathbf{y},\bar{\theta})\mathrm{d}\mathbf{z} \approx \frac{1}{M}\sum_{m=1}^{M} \ln p\left(\mathbf{y},\mathbf{z}^{(m)}|\bar{\theta}\right),$$

where $\{\mathbf{z}^{(m)}\}_{m=1}^{M}$ are drawn from the posterior distribution $p(\mathbf{z}|\mathbf{y},\bar{\theta})$.

For the second term in (22), if $p(\mathbf{z}|\mathbf{y},\bar{\theta})$ is a standard distribution, $\mathcal{H}(\bar{\theta}|\bar{\theta})$ can be easily evaluated from MCMC output as

$$\mathcal{H}(\bar{\theta}|\bar{\theta}) = \int \ln p(\mathbf{z}|\mathbf{y},\bar{\theta})p(\mathbf{z}|\mathbf{y},\bar{\theta})\mathrm{d}\mathbf{z} \approx \frac{1}{M}\sum_{m=1}^{M} \ln p\left(\mathbf{z}^{(m)}|\mathbf{y},\bar{\theta}\right).$$

However, if $p(\mathbf{z}|\mathbf{y},\bar{\theta})$ is not a standard distribution, an alternative approach has to be used, depending on the specific model in consideration. We now consider two situations.

First, if the complete-data $(\mathbf{y}_i,\mathbf{z}_i)$ are independent when $i \neq j$, and $\mathbf{z}_i$ is low-dimensional, say ≤ 5, then a nonparametric approach may be used to approximate $p(\mathbf{z}|\mathbf{y},\theta)$. Note that

$$\mathcal{H}(\theta|\theta) = \int \ln p(\mathbf{z}|\mathbf{y},\theta)\pi(\mathbf{z}|\mathbf{y},\theta)\mathrm{d}\mathbf{z} = \sum_{i=1}^{n}\int \ln p(\mathbf{z}_i|\mathbf{y}_i,\theta)\pi(\mathbf{z}_i|\mathbf{y},\theta)\mathrm{d}\mathbf{z}_i = \sum_{i=1}^{n}\mathcal{H}_i(\theta|\theta).$$

Computation of $\mathcal{H}_i(\theta|\theta)$ requires an analytic approximation to $p(\mathbf{z}_i|\mathbf{y}_i,\theta)$ via a nonparametric method. In particular, MCMC allows one to draw some effective samples from $p\left(\mathbf{z}_i|\mathbf{y}_i,\theta\right)$.

Using these random samples, one can then use nonparametric techniques such as the kernel-based methods to approximate $p(\mathbf{z}_i|\mathbf{y}_i, \theta)$. In a recent study, Ibrahim et al. (2008) suggested using a truncated Hermite expansion to approximate $p(\mathbf{z}_i|\mathbf{y}_i, \theta)$.

As a simple illustration, we apply this method to the Clark model. When the Gaussian kernel method is used, we get $\ln p(\mathbf{y}|\bar{\theta}) = -1448.97$, $\text{DIC}_L = 2901.46$ for Model 1 and $\ln p(\mathbf{y}|\bar{\theta}) = -1449.41$, $\text{DIC}_L = 2902.42$ for Model 2. These two sets of numbers are nearly identical. However, if the latent variable models are regarded as parameters, we get $\text{DIC}_7 = 2884.37$ for Model 1 and $\text{DIC}_7 = 2852.85$ for Model 2. The highly distinctive difference between them suggests that DIC_7 is not a reliable model selection criterion for the model. Note that DIC_1 is very difficult to compute in this case.

Second, for some latent variable models, the latent variables $\mathbf{z}$ follow a multivariate normal distribution, and the observed variables $\mathbf{y}$ are independent conditional on $\mathbf{z}$. This class of models is referred to as the Gaussian latent variable models in the literature. In economics and finance, many latent variable models belong to this class of models, including dynamic linear models, dynamic factor models, various forms of stochastic volatility models, and credit risk models. In these models, the observed-data likelihood is non-Gaussian but has a Gaussian flavor in the sense that the posterior distribution, $p(\mathbf{z}|\mathbf{y}, \theta)$, may be expressed as,

$$p(\mathbf{z}|\mathbf{y}, \theta) \propto \exp\left(-\frac{1}{2}\mathbf{z}'V(\theta)\mathbf{z} + \sum_{i=1}^{n} \ln p\left(\mathbf{y}_i|\mathbf{z}_i, \theta\right)\right).$$

Rue et al. (2004) and Rue et al. (2009) showed that this type of posterior distribution can be well approximated by a Gaussian distribution via the Laplace approximation, that is,

$$p(\mathbf{z}|\mathbf{y}, \theta) \propto \exp\left(-\frac{1}{2}\mathbf{z}'(V(\theta) + diag(\mathbf{c}))\mathbf{z}\right),$$

where $\mathbf{c}$ comes from the second-order term in the Taylor expansion of $\sum_{i=1}^{n} \ln p(\mathbf{y}_i|\mathbf{z}_i)$ at the mode of $p(\mathbf{z}|\mathbf{y}, \theta)$. The Laplace approximation may be employed to compute $\mathcal{H}(\bar{\theta}|\bar{\theta})$. After $p(\mathbf{y}|\bar{\theta})$ is obtained, it is easy to obtain $D(\bar{\theta})$. It is important to point out that the numerical evaluation of $p(\mathbf{y}|\bar{\theta})$ is needed only once, that is, at the posterior mean.

To compute P_L, we have to calculate the second derivative of the observed-data likelihood function in P_L. Under the mild regularity condition, Louis (1982) showed that this second derivative may be expressed as:

$$\begin{aligned} \mathbf{I}(\theta) &= -\frac{\partial^2 \mathcal{L}_o(\mathbf{y}|\theta)}{\partial\theta\partial\theta'} = E_{\mathbf{z}|\mathbf{y},\theta}\left\{-\frac{\partial^2 \mathcal{L}_c(\mathbf{x}|\theta)}{\partial\theta\partial\theta'}\right\} - Var_{\mathbf{z}|\mathbf{y},\theta}\left\{S(\mathbf{x}|\theta)\right\} \qquad (23) \\ &= E_{\mathbf{z}|\mathbf{y},\theta}\left\{-\frac{\partial^2 \mathcal{L}_c(\mathbf{x}|\theta)}{\partial\theta\partial\theta'} - S(\mathbf{x}|\theta)S(\mathbf{x}|\theta)'\right\} + E_{\mathbf{z}|\mathbf{y},\theta}\{S(\mathbf{x}|\theta)\}E_{\mathbf{z}|\mathbf{y},\theta}\{S(\mathbf{x}|\theta)\}', \end{aligned}$$

where $S(\mathbf{x}|\theta) = \partial\mathcal{L}_c(\mathbf{x}|\theta)/\partial\theta$ and all the expectations are taken with respect to the conditional distribution of $\mathbf{z}$ given $\mathbf{y}$ and θ.

If $\mathcal{Q}$ function has an analytical expression, Oakes (1999) showed that the second derivative has an equivalent expression

$$\mathbf{I}(\theta) = -\frac{\partial^2 \mathcal{L}_o(\mathbf{y}|\theta)}{\partial\theta\partial\theta'} = \left\{ -\frac{\partial^2 \mathcal{Q}(\theta|\theta^*)}{\partial\theta\partial\theta'} - \frac{\partial^2 \mathcal{Q}(\theta|\theta^*)}{\partial\theta\partial\theta^{*\prime}} \right\}_{\theta^*=\theta}. \tag{24}$$

If the analytical $\mathcal{Q}$ function not available, we may approximate the second derivatives by,

$$\begin{aligned} & E_{\mathbf{z}|\mathbf{y},\theta}\left\{ -\frac{\partial^2 \mathcal{L}_c(\mathbf{x}|\theta)}{\partial\theta\partial\theta'} - S(\mathbf{x}|\theta)S(\mathbf{x}|\theta)' \right\}, \\ \approx & -\frac{1}{M}\sum_{m=1}^{M}\left\{ \frac{\partial^2 \mathcal{L}_c(\mathbf{y},\mathbf{z}^{(m)}|\theta)}{\partial\theta\partial\theta'} + S(\mathbf{y},\mathbf{z}^{(m)}|\theta)S(\mathbf{y},\mathbf{z}^{(m)}|\theta)' \right\}, \\ & E_{\mathbf{z}|\mathbf{y},\theta}\{S(\mathbf{x}|\theta)\} \approx \frac{1}{M}\sum_{m=1}^{M} S(\mathbf{y},\mathbf{z}^{(m)}|\theta), \end{aligned}$$

where $\{\mathbf{z}^{(m)}, m = 1, 2, \cdots, M\}$ are random observations drawn from $p(\mathbf{z}|\mathbf{y}, \theta)$.

Although EM algorithm is a very general approach to analyzing latent variable models, it is very cumbersome to deal with dynamic latent variable models, such as state-space models, because we have to compute the derivatives recursively (Doucet and Shephard, 2012). Alternatively, one can compute DIC_L using the Kalman filter and particle filters.

3.4.2 Computing DIC_L by the Kalman filter

In economics, many time series models can be represented by a linear Gaussian state-space form. The Kalman filter is an efficient recursive method for computing the optimal linear forecasts in such models. It also gives the exact likelihood function of the model. Here, we only present the basic idea of the Kalman filter for analyzing linear state-space models. One may refer to Harvey (1989) for the detailed textbook treatment.

Consider a general linear state-space model,

$$z_t = Tz_{t-1} + R\varepsilon_t, y_t = D + Cz_t + \xi_t,$$

where $\varepsilon_t \sim N(0, Q)$, $\xi_t \sim N(0, H)$, T is $n_s \times n_s$, R is $n_s \times n_e$, D is $n \times 1$, C is $n \times n_s$, Q is $n_e \times n_e$, H is $n \times n$. These six coefficient matrices are functions of a vector of parameters θ which is $n_q \times 1$.

Let $z_t^s = E(z_t|\mathbf{y}^s)$, $\Sigma_t^s = E\{(z_t - z_t^s)(z_t - z_t^s)'|\mathbf{y}^s\}$. With the initial conditions, z_0^0 and Σ_0^0, for $t = 1, 2, \ldots, n$, the Kalman filter recursively implements the following steps

$$z_t^{t-1} = Tz_{t-1}^{t-1}, \Sigma_t^{t-1} = T\Sigma_{t-1}^{t-1}T' + RQR',$$

and

$$z_t^t = z_t^{t-1} + K_t\left(y_t - D - Cz_t^{t-1}\right), \Sigma_t^t = \left[I_{n_s} - K_tC\right]\Sigma_t^{t-1},$$

where

$$K_t = \Sigma_t^{t-1} C' \left[C \Sigma_t^{t-1} C' + H \right]^{-1}.$$

The observed-data log-likelihood is given by

$$\begin{aligned} \ln p(\mathbf{y}|\theta) &= -\sum_{t=1}^{n} \left[\frac{n}{2} \ln 2\pi + \frac{1}{2} \ln |F_t| + \frac{1}{2} \left(y_t - D - C z_t^{t-1} \right)' F_t^{-1} \left(y_t - D - C z_t^{t-1} \right) \right] \\ &= -\sum_{t=1}^{n} \left[\frac{n}{2} \ln 2\pi + \frac{1}{2} \ln |F_t| + \frac{1}{2} \omega_t' F_t^{-1} \omega_t \right], \end{aligned}$$

where $F_t = C P_t^{t-1} C' + H$, $\omega_t = y_t - D - C z_t^{t-1}$. Clearly, $\ln p(y|\theta)$ has to be calculated recursively since F_t and z_t^{t-1} are only available recursively. Similarly, $s_t(\theta)$ and $h_t(\theta)$ has to be computed recursively. To calculate $s_t(\theta)$ and $h_t(\theta)$, we need to calculate the first and second-order derivatives of $|F_t|$, $\omega_t' F_t^{-1} \omega_t$ recursively. For details, one can refer to Iskrev (2008).

3.4.3 Computing DIC_L by particle filters

In practice, the nonlinear non-Gaussian state-space models have been widely used in empirical works, but they cannot be analyzed using the Kalman filter. Instead, one can use another class of recursive filtering algorithms known as particle filters. We only present the basic idea of particle filters here and refer the reader to recent review papers on particle filters by Doucet and Johansen (2009) and Creal (2012) for greater details.

Let $z_{t+1}|z_t \sim f(z_{t+1}|z_t, \theta)$ and $y_t|z_t \sim g(y_t|z_t, \theta)$. Let the initial density of z be $\mu(z|\theta)$. The joint density of $\left(\mathbf{z}^t, \mathbf{y}^t\right)$ is

$$p\left(\mathbf{z}^t, \mathbf{y}^t|\theta\right) = \mu(z_1|\theta) \prod_{k=2}^{t} f(z_k|z_{k-1}, \theta) \prod_{k=1}^{t} g(y_k|z_k, \theta),$$

and hence,

$$p\left(\mathbf{y}^t|\theta\right) = \int p\left(\mathbf{z}^t, \mathbf{y}^t|\theta\right) \mathrm{d}\mathbf{z}^t.$$

For nonlinear non-Gaussian state-space models, neither $p\left(\mathbf{z}^t|\mathbf{y}^t, \theta\right)$ nor $p\left(\mathbf{y}^t|\theta\right)$ are available in closed-form. The goal here is to calculate $p\left(\mathbf{z}^t|\mathbf{y}^t, \theta\right)$, $p\left(\mathbf{y}^t|\theta\right)$, and $\mathbf{s}(\mathbf{y}^t, \theta)$ sequentially for $t = 1, \ldots, n$. The idea of the using particle filters is to approximate $p\left(\mathbf{z}^t|\mathbf{y}^t, \theta\right)\mathrm{d}\mathbf{z}^t$ by its empirical measure. An example of particle filters is the Sequential Important Sampling and Resampling (SISR) algorithm which iterates the following step for $i = 1, \ldots, N$,

Step 1: At $t = 1$, $z_1^{(i)} \sim \mu(\cdot)$,

$$w_1\left(\mathbf{z}^{1(i)}\right) = \frac{\mu\left(z_1^{(i)}|\theta\right) g\left(y_1|z_1^{(i)}, \theta\right)}{q_1\left(z_1^{(i)}\right)}, \quad W_1^{(i)} = \frac{w_1\left(\mathbf{z}^{1(i)}\right)}{\sum_{i=1}^{N} w_1\left(\mathbf{z}^{1(i)}\right)},$$

$\mathbf{z}^{1(i)} = z_1^{(i)}$. Resample $\left(W_1^{(i)}, \mathbf{z}^{1(i)}\right)$ to obtain new particles $\left(\frac{1}{N}, \widetilde{\mathbf{z}}^{1(i)}\right)$.

Step 2: At $t \geq 2$, $z_t^{(i)} \sim q_n\left(\cdot|\widetilde{\mathbf{z}}^{t-1(i)}\right)$,

$$w_t\left(\mathbf{z}^{t(i)}\right) = \frac{f\left(z_t^{(i)}|\widetilde{z}_{t-1}^{(i)}, \theta\right) g\left(y_t|\widetilde{z}_t^{(i)}, \theta\right)}{q_t\left(z_t^{(i)}|\widetilde{\mathbf{z}}^{t-1(i)}\right)}, \quad W_t^{(i)} = \frac{w_t\left(\mathbf{z}^{t(i)}\right)}{\sum_{i=1}^N w_t\left(\mathbf{z}^{t(i)}\right)},$$

$\mathbf{z}^{t(i)} = \left(\widetilde{\mathbf{z}}^{t-1(i)}, z_t^{(i)}\right)$. Resample $\left(W_t^{(i)}, \mathbf{z}^{t(i)}\right)$ to obtain new particles $\left(\frac{1}{N}, \widetilde{\mathbf{z}}^{t(i)}\right)$.

Step 3: Approximate the conditional distribution $p_\theta\left(\mathrm{d}\mathbf{z}^t|\mathbf{y}^t, \theta\right)$ by its empirical measure

$$\hat{p}\left(\mathrm{d}\mathbf{z}^t|\mathbf{y}^t, \theta\right) = \sum_{i=1}^N W_t^{(i)} \delta_{\mathbf{z}^{t(i)}}\left(\mathrm{d}\mathbf{z}^t\right) \quad \text{or} \quad \widetilde{p}_\theta\left(d\mathbf{z}^t|\mathbf{y}^t, \theta\right) = \frac{1}{N}\sum_{i=1}^N \delta_{\widetilde{\mathbf{z}}^{t(i)}}\left(\mathrm{d}\mathbf{z}^t\right),$$

and

$$\hat{p}\left(y_t|\mathbf{y}^{t-1}, \theta\right) = \frac{1}{N}\sum_{i=1}^N w_t\left(\mathbf{z}^{t(i)}\right),$$

where N is the number of particles and $q_t\left(\cdot|\cdot\right)$ is the proposal density.

With the empirical measure $\left\{\hat{p}\left(\mathrm{d}\mathbf{z}^t|\mathbf{y}^t, \theta\right)\right\}_{t=1:n}$, we can approximate the integral

$$I_t = \int \varphi_t\left(\mathbf{z}^t\right) p\left(\mathbf{z}^t|\mathbf{y}^t, \theta\right) \mathrm{d}\mathbf{z}^t,$$

by

$$\hat{I}_t = \int \varphi_t\left(\mathbf{z}^t\right) \hat{p}\left(\mathrm{d}\mathbf{z}^t|\mathbf{y}^t, \theta\right) = \sum_{i=1}^N W_t^{(i)} \varphi_t\left(\mathbf{z}^{t(i)}\right),$$

for $t = 1, \cdots, n$, where $\varphi_t\left(\mathbf{z}^t\right)$ is the target function. If $\varphi_t\left(\mathbf{z}^t\right) = \partial \ln p\left(\mathbf{z}^t, \mathbf{y}^t|\theta\right)/\partial\theta$, then

$$\mathbf{s}(\mathbf{y}^t, \theta) = \int \frac{\partial \ln p\left(\mathbf{z}_t, \mathbf{y}^t|\theta\right)}{\partial\theta} p\left(\mathbf{z}_t|\mathbf{y}^t, \theta\right) \mathrm{d}\mathbf{z}_t, \; -\mathbf{H}(\mathbf{y}^t, \theta) = \mathbf{s}(\mathbf{y}^t, \theta)\mathbf{s}(\mathbf{y}^t, \theta)' - \frac{\partial^2 p\left(\mathbf{y}^t|\theta\right)/\partial\theta\partial\theta'}{p\left(\mathbf{y}^t|\theta\right)}$$

where

$$\begin{aligned}\frac{\partial^2 p\left(\mathbf{y}^t|\theta\right)/\partial\theta\partial\theta'}{p\left(\mathbf{y}^t|\theta\right)} &= \int \frac{\partial \ln p\left(\mathbf{z}_t, \mathbf{y}^t|\theta\right)}{\partial\theta} \frac{\partial \ln p\left(\mathbf{z}_t, \mathbf{y}^t|\theta\right)}{\partial\theta}' p\left(\mathbf{z}_t|\mathbf{y}^t, \theta\right) \mathrm{d}\mathbf{z}_t \\ &\quad + \int \frac{\partial^2 \ln p\left(\mathbf{z}_t, \mathbf{y}^t|\theta\right)}{\partial\theta\partial\theta'} p\left(\mathbf{z}_t|\mathbf{y}^t, \theta\right) \mathrm{d}\mathbf{z}_t,\end{aligned}$$

by the Fisher and Louis identities that are based only on the marginal density $p\left(\mathbf{z}_t|\mathbf{y}^t, \theta\right)$ (Poyiadjis et al, 2011). Therefore, $s(y^t, \theta)$ and $H(y^t, \theta)$ can be obtained recursively.

Based on different proposal densities $q_t\left(\cdot|\cdot\right)$, different particle filtering algorithms have been proposed in the literature, including the bootstrap particle filters of Gordon et al. (1993) and the auxiliary particle filters of Pitt and Shephard (1999). In this paper, we use the auxiliary particle filters to compute $s(y^t, \theta)$, $H(y^t, \theta)$. The details about how to compute them via particle filters can be found in Poyiadjis, et al (2011) and Doucet and Shephard (2012).

4 DIC for Misspecified Models

According to Assumption 9, DIC_L requires all candidate models be good approximations to DGP. The same requirement is needed for AIC and DIC_1. In most applications, however, this assumption is too strong. Quoting Box (1976), "all models are wrong, but some are useful." In this section, following a referee's suggestion, we relax this assumption and introduce a new DIC (namely DIC_M) to compare misspecified models, namely, when all candidate models violate Assumption 9. We first develop DIC_M and obtain its asymptotic properties. Following a suggestion of another referee, we then discuss BFs and BIC in the context of misspecified models. Finally, we design a simple simulation study to compare the performance of alternative model selection criteria.

4.1 DIC_M for misspecified models

The asymptotic justification of AIC and DIC_1 requires all candidate models be correctly specified or good approximations to the DGP. If a candidate model is misspecified, the expected KL divergence between the DGP and $p\left(\mathbf{y}_{rep}|\hat{\theta}(\mathbf{y})\right)$ can be expressed as

$$\begin{aligned} & E_{\mathbf{y}}\left\{2\times KL\left[g\left(\mathbf{y}_{rep}\right),p\left(\mathbf{y}_{rep}|\hat{\theta}(\mathbf{y})\right)\right]\right\}=2C+E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\hat{\theta}(\mathbf{y})\right)\right] \\ = \;& 2C+E_{\mathbf{y}}\left\{-2\ln p\left(\mathbf{y}|\hat{\theta}(\mathbf{y})\right)-2\mathbf{tr}\left\{\mathbf{B}_n\left(\theta_n^p\right)\mathbf{H}_n^{-1}\left(\theta_n^p\right)\right\}\right\}+o(1), \end{aligned} \tag{25}$$

where $\hat{\theta}(\mathbf{y})$ denotes the MLE of θ in the misspecified model. As before, we write $\hat{\theta}(\mathbf{y})$ as $\hat{\theta}$. Note the difference between (25) and (4) for AIC. Based on (25), TIC is defined as

$$\text{TIC}=-2\ln p\left(\mathbf{y}|\hat{\theta}\right)+2P_T, \tag{26}$$

where P_T is a consistent estimator of $-\mathbf{tr}\left\{\mathbf{B}_n\left(\theta_n^p\right)\mathbf{H}_n^{-1}\left(\theta_n^p\right)\right\}$. TIC is an asymptotically unbiased estimator of the expected KL divergence minus $2C$ when a candidate model is misspecified. Equation (26) was first proposed by Takeuchi (1976) for independent data. Stone (1977) derived the same results from the viewpoint of cross-validation. Clearly, finding a consistent estimator for $-\mathbf{tr}\left\{\mathbf{B}_n\left(\theta_n^p\right)\mathbf{H}_n^{-1}\left(\theta_n^p\right)\right\}$ is critical to TIC.

Under Assumptions 1-8, $\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)$ is a consistent estimator for $\mathbf{H}_n^{-1}\left(\theta_n^p\right)$, that is,

$$\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)-\mathbf{H}_n^{-1}\left(\theta_n^p\right)\xrightarrow{p}0. \tag{27}$$

Newey and West (1987) proposed a heteroskedasticity and autocorrelation consistent (HAC) estimator of $\mathbf{B}_n\left(\theta_n^p\right)$ defined by

$$\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)=\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\mathbf{s}_t\left(\hat{\theta}\right)\mathbf{s}_\tau\left(\hat{\theta}\right)'k\left(\frac{t-\tau}{\gamma_n}\right),$$

where $k(\cdot)$ is a kernel function and γ_n is the bandwidth. The penalty term P_T then becomes

$$P_T = -\mathbf{tr}\left\{\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right\}. \tag{28}$$

To ensure consistency and positive semidefiniteness of $\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)$, following de Jong and Davidson (2000), we add three more assumptions. The first two are about the kernel function and the bandwidth parameter, while the last one is about the score function $\mathbf{s}_t\left(\theta_n^p\right)$.

Assumption 11: Assume the kernel function $k(\cdot) \in \mathcal{H}$, where

$$\mathcal{H} = \left\{ \begin{array}{c} k(\cdot): R \to [-1,1], k(x) = k(-x), \text{for any } x \in R, \\ \int_{-\infty}^{+\infty} |k(x)|\, dx < \infty,\ \int_{-\infty}^{+\infty} \psi(\xi)\, d\xi < \infty, \\ k(\cdot) \text{ is continuous at 0 and at all but a finite number of points in } R \end{array} \right\},$$

where

$$\psi(\xi) = (2\pi)^{-1} \int_{-\infty}^{+\infty} k(x)\, e^{i\xi x} dx.$$

Assumption 12: The bandwidth parameter γ_n is an increasing function of sample size n and $\gamma_n = o\left(n^{1/2}\right)$.

Assumption 13: The expectation of the score function $E\left(\mathbf{s}_t\left(\theta_n^p\right)\right) = 0$ for any t.

Remark 4.1 *In Assumption 11, the function class $\mathcal{H}$ includes many well-known kernel functions, such as Bartlett, Parzen, Quadratic Spectral, and Tukey-Hanning kernels. It ensures that $\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)$ is positive semidefinite with probability 1; see Andrews (1991). Note that $\mathcal{H}$ does not include truncated kernels. If Assumption 9 is satisfied, $P_T = P + o_p(1)$.*

Remark 4.2 *From Assumption 1-8, we have $\sqrt{n}\left(\hat{\theta} - \theta_n^p\right) = O_p(1)$; see Gallant and White (1988). In the online supplement, we show that our Assumptions 1-8 and 11-13 imply the set of regularity conditions of de Jong and Davidson (2000) which in turn implies that*

$$\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right) - \mathbf{B}_n\left(\theta_n^p\right) \xrightarrow{p} 0. \tag{29}$$

Our assumptions are more primitive than those imposed by de Jong and Davidson (2000). In the same online supplement, we also show that if Assumption 13 does not hold, it may not be true that $\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right) - \mathbf{B}_n\left(\theta_n^p\right) \xrightarrow{p} 0$. Together with (27), (29) implies that

$$P_T - \mathbf{tr}\left\{\mathbf{B}_n\left(\theta_n^p\right)\mathbf{H}_n^{-1}\left(\theta_n^p\right)\right\} = \mathbf{tr}\left\{\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right\} - \mathbf{tr}\left\{\mathbf{B}_n\left(\theta_n^p\right)\mathbf{H}_n^{-1}\left(\theta_n^p\right)\right\} \xrightarrow{p} 0. \tag{30}$$

Hence, the frequentist asymptotic justification of TIC is provided under misspecified models, and the asymptotic justification of TIC requires Assumption 13.

Clearly, TIC requires the MLE of θ be available in the misspecified model. If only MCMC samples for θ are available, a model selection criterion based on $\bar{\theta}$ is needed. We propose the following DIC to compare misspecified models,

$$\mathrm{DIC}_M = D\left(\bar{\theta}\right) + 2P_M \quad \text{with } P_M = \mathbf{tr}\left\{n\bar{\mathbf{\Omega}}_n\left(\bar{\theta}\right)V\left(\bar{\theta}\right)\right\}, \tag{31}$$

where $V\left(\overline{\theta}\right)$ is the posterior covariance matrix given by $V\left(\overline{\theta}\right) = E\left[\left(\theta-\overline{\theta}\right)\left(\theta-\overline{\theta}\right)'|\mathbf{y}\right]$ which, when multiplied by $-n$, consistently estimate $\mathbf{H}_n^{-1}\left(\theta_n^p\right)$ according to Lemma 3.2. From Li et al. (2017), we have $D\left(\overline{\theta}\right) = D\left(\hat{\theta}\right) + O_p\left(1/n\right)$.[4] So the only thing that remains to be verified is $\bar{\mathbf{\Omega}}_n\left(\overline{\theta}\right) - \mathbf{B}_n\left(\theta_n^p\right) \xrightarrow{p} 0$.

Theorem 4.1 *Under Assumptions 1-8 and 10-12, we have*

$$\bar{\mathbf{\Omega}}_n\left(\overline{\theta}\right) - \bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right) \xrightarrow{p} 0, \tag{32}$$

$$P_M = P_T + \frac{\gamma_n}{n}C_1^M + \frac{1}{n}C_2^M + O_p\left(\frac{\gamma_n}{n^2}\right), \tag{33}$$

$$DIC_M = TIC + \frac{\gamma_n}{n}D_1^M + \frac{1}{n}D_2^M + O_p\left(\frac{\gamma_n}{n^2}\right), \tag{34}$$

where γ_n is defined in Assumption 12 and

$$\begin{aligned} C_1^M &= vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)'\tilde{U}_1\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\frac{\hat{p}^{(1)}}{\hat{p}} \\ &\quad -\frac{1}{2}vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)'\tilde{U}_1\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right), \end{aligned}$$

$$\begin{aligned} C_2^M &= -\frac{1}{2n}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\otimes vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\hat{\theta}\right)\right] \\ &\quad +\frac{1}{2n}\mathbf{tr}\left[\begin{array}{c}\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)' \\ \times\left[\left(vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)'\otimes\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right]\end{array}\right] \\ &\quad +\frac{1}{2n}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\otimes\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right] \\ &\quad -\frac{1}{n}\mathbf{tr}\left[\left[\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\frac{\hat{p}^{(1)}}{\hat{p}}\right)'\otimes\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\right] \\ &\quad +\frac{1}{n}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\frac{\hat{p}^{(2)}}{\hat{p}}\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right] - \frac{1}{n}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\frac{\hat{p}^{(1)}}{\hat{p}}\frac{\hat{p}^{(1)\prime}}{\hat{p}}\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right], \end{aligned}$$

$$\tilde{U}_1 = \frac{1}{n\gamma_n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\left[l_\tau^{(2)}\left(\hat{\theta}\right)\otimes\mathbf{s}_t\left(\hat{\theta}\right) + \mathbf{s}_\tau\left(\hat{\theta}\right)\otimes l_t^{(2)}\left(\hat{\theta}\right)\right]k\left(\frac{t-\tau}{\gamma_n}\right),$$

$$D_1^M = 2C_1^M, D_2^M = C_{21} - C_{23} - \frac{1}{4}C_{12} + 2C_2^M.$$

Remark 4.3 *According to Theorem 4.1, under Assumptions 1-8 and 10-12, DIC_M and TIC are asymptotically equivalent. Thus, DIC_M can be regarded as a Bayesian version of TIC. If,*

[4] While Li et al (2017) assumes correct model specification, such an assumption is not needed to obtain the relationship between $D\left(\overline{\theta}\right)$ and $D\left(\hat{\theta}\right)$.

in addition, Assumption 13 holds, then both (29) and (30) hold, justifying TIC asymptotically. The same frequentist justification applies to DIC_M due to (32). Therefore, DIC_M and TIC provide asymptotically unbiased estimation to the corresponding expected KL divergence.

Remark 4.4 *For misspecified latent variable models, if DIC_M is calculated based on $p(\mathbf{y}|\theta)$ not on $p(\mathbf{y}|\theta, \mathbf{z})$, the frequentist asymptotic justification of DIC_M is also applicable.*

Remark 4.5 *Since DIC_M applies to both correctly specified and misspecified models while DIC_L applies only to asymptotically correctly specified models, it may be attempting to use DIC_M rather than DIC_L to select a model. However, DIC_M requires the Fisher information matrix, while is usually easier to compute than the Hessian information matrix required by DIC_L. Hence, if a candidate model that is "locally" misspecified in the sense of Assumption 9 and the empirical Fisher information matrix is too difficult to evaluate or numerically unstable, DIC_L is preferable. This comparison applies to AIC and TIC, which may helps explain why AIC is used more widely than TIC in practice.*

4.2 BF and BIC

There are two strands of literature on model selection. The first strand aims to answer the following question: which model give the best prediction of out-of-sample observations generated by the same mechanism that gives rise to the observed data? Clearly this is a utility-based approach where the utility is prediction. Based on hypothetically replicate data generated by the same mechanism that gives rise to the observed data, some predictive information criteria have been proposed for model comparison. These criteria minimize an expected loss function associated with the prediction. AIC, TIC, DIC, DIC_L, and DIC_M all belong to this strand.

The second strand aims to answer the following question: which model best explains the observed data? The BF and BIC belong to this strand. They compare competing models by examining model posterior probabilities and search for the "true" model. A recent development of the BF in economics is found in Inoue and Shintani (2018). BIC is a large sample approximation to the log-marginal likelihood, although it is based on the MLE. Many applications of BIC in economics can be found. Both BFs and BIC enjoy the property of consistency, that is, when the true DGP is one of the candidate models, BFs and BIC select it with probability approaching 1 when the sample size goes to infinity. For more information about different model selection criteria, see Burnham and Anderson (2002) and Vehtari and Ojanen (2012).

In the Bayesian framework, the BF is arguably the most widely used statistic for model comparison. Suppose there are two candidate models, M_1 and M_2. The BF of M_1 against M_2 is defined as

$$B_{12} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}, \tag{35}$$

where $p(y|M_k)$ is the marginal likelihood of model M_k which is obtained by

$$p(\mathbf{y}|M_k) = \int_{\mathbf{\Theta}_k} p(\mathbf{y}|\theta_k, M_k)p(\theta_k|M_k)\mathrm{d}\theta_k, \quad \theta_k \in \mathbf{\Theta}_k, k = 1, 2,$$

where θ_k is the set of parameters in M_k, $p(\mathbf{y}|\theta_k, M_k)$ the likelihood function of M_k, $p(\theta_k|M_k)$ the prior of θ_k in M_k. If $B_{12} > 1$, M_1 is preferred to M_2 and vice versa.

Remark 4.6 *In practice, the BF is subject to several problems. First, it is not well-defined with improper priors. Second, calculation of BFs requires comparing marginal likelihoods. When the dimension of parameter space is large, as is typical in latent variable models, high-dimensional integrations pose a formidable computational challenge. Third, it is well-known that the BF suffers from the Jeffreys-Lindley paradox when a vague and proper prior is employed; see Kass and Raftery (1995).*

Based on the Laplace approximation, Schwarz (1978) showed that the log-marginal likelihood can be approximated by

$$\ln p\left(\mathbf{y}|M_k\right) = \ln p\left(\mathbf{y}|\hat{\theta}_k, M_k\right) + \ln p\left(\hat{\theta}_k|M_k\right) + \frac{P_k\pi}{2} - \frac{P_k \ln n}{2} - \frac{\left|-\bar{\mathbf{H}}_n\left(\hat{\theta}_k\right)\right|}{2} + O_p\left(\frac{1}{n}\right), \quad (36)$$

where $\hat{\theta}_k$ is the MLE of θ_k and $\bar{\mathbf{H}}_n\left(\hat{\theta}_k\right) = \frac{1}{n}\frac{\partial^2 \ln p\left(\mathbf{y}|\hat{\theta}_k, M_k\right)}{\partial\theta_k\partial\theta_k'}$, and P_k is the dimension of θ_k. Ignoring all the $O_p(1)$ terms in (36) and under noninformative priors such as $p\left(\theta_k|M_k\right) \propto 1$, Schwarz defined BIC_k as

$$\mathrm{BIC}_k := -2\ln p\left(\mathbf{y}|\hat{\theta}_k, M_k\right) + P_k \ln n,$$

where, as in AIC and TIC, $-2\ln p\left(\mathbf{y}|\hat{\theta}_k, M_k\right)$ is used to measure the model fit, but $P_k \ln n$ is the new penalty term. Obviously, BIC_k provides an approximation of $-2\ln\left(\mathbf{y}|M_k\right)$.

Remark 4.7 *From the theoretical viewpoint, different criteria have different theoretical properties. BIC and BFs are consistent if the true model is one of the candidate models while AIC, TIC, DIC_L, and DIC_M aim to provide the asymptotically unbiased estimator of the expected KL divergence between the DGP and a predictive distribution. When the true model is not included as a candidate model, which is often the case in practice, it is not clear what the best model selected by BIC and BFs can achieve. In this case, if one is concerned with the KL divergence between the DGP and a predictive distribution, it is expected that TIC and DIC_M perform better than BIC and BFs. Moreover, when the sample size is small, even when the true model is a candidate model, BIC and BFs may not select the true model. Again, if one is concerned with the KL divergence between the DGP and a predictive distribution, AIC and DIC_L can perform better than BIC and BFs.*

4.3 A simulation study

In this subsection, we design a simple experiment to compare alternative model selection criteria when the true DGP is not included into the set of candidate models. In other words, all candidate models are misspecified.

Following Ding et al. (2019), we generate data from the following model

$$y_i = \ln(1 + 46x_i) + e_i,\ e_i \sim N(0,1), i = 1, \ldots, n, \tag{37}$$

where $x_i = 0.7(i-1)/n$ which is fixed under repeated sampling by design. In practice, researchers do not know the functional form. Suppose the following set of polynomial regressions is considered,

$$M_k : y_i = \sum_{j=0}^{k-1} \beta_{k,j+1} x_i^j + u_i, \tag{38}$$

where $k = 1, \ldots, \lfloor n^{1/3} \rfloor$ and u_i is assumed to be $N(0, \sigma^2)$. When $k \to \infty$ as $n \to \infty$, the polynomial regression is related to the sieve estimator which uses progressively more complex models to estimate an unknown function as more data becomes available. In our experiment, we estimate and compare all the candidate models $\left\{M_k, k = 1, \ldots, \lfloor n^{1/3} \rfloor\right\}$. In M_k, $\sum_{j=0}^{k-1} \beta_{k,j+1} x_i^j$ is used to approximate $\ln(1 + 46x_i)$. Let $\beta_k = (\beta_1, \ldots, \beta_k)'$ so that $\theta_k = \left(\beta_k', \sigma^2\right)$ and the number of parameters is $k + 1$. Let $\mathbf{x}^j = \left(x_1^j, x_2^j, \ldots, x_n^j\right)'$, $\mathbf{X}_k = \left(\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^{k-1}\right)$, and $\mathbf{X} = \left(\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^{[n^{1/3}]-1}\right)$.

Two different sample sizes are considered, $n = 100, 500$. For each candidate model M_k, we obtain the MLE of θ_k, denoted by $\widehat{\theta}_k = (\hat{\beta}_k', \hat{\sigma}^2)$, and then calculate AIC, TIC, and BIC. $\widehat{\theta}_k$, which is also the least squares estimate, has a closed-form expression for this model.

The following g-prior is used for θ_k when we conduct the Bayesian analysis,

$$\pi\left(\sigma^2\right) \propto \frac{1}{\sigma^2}, \quad \beta_k \sim N\left(\beta_{k,0}, g\sigma^2 \left(\mathbf{X}_k'\mathbf{X}_k\right)^{-1}\right), \tag{39}$$

where $g = n$ denotes the unit information prior (Kass and Wasserman, 1995) in the normal regression case. The posterior mean and the posterior variance of θ_k are

$$E(\beta_k|\mathbf{y}, \mathbf{X}) = \frac{g}{g+1}\left(\frac{\beta_{k,0}}{g} + \hat{\beta}_k\right), \tag{40}$$

$$E\left(\sigma^2|\mathbf{y}, \mathbf{X}\right) = \frac{s^2 + \frac{1}{g+1}\left(\hat{\beta}_k - \beta_{k,0}\right)' \mathbf{X}_k'\mathbf{X}_k \left(\hat{\beta}_k - \beta_{k,0}\right)}{n-2}, \tag{41}$$

$$Var(\beta_k|\mathbf{y}, \mathbf{X}) = \frac{g}{g+1}\left(\mathbf{X}_k'\mathbf{X}_k\right)^{-1} E\left(\sigma^2|\mathbf{y}, \mathbf{X}\right), \tag{42}$$

$$Var(\sigma^2|\mathbf{y}, \mathbf{X}) = \frac{2E\left(\sigma^2|\mathbf{y}, \mathbf{X}\right)^2}{n-4}, \tag{43}$$

$$Cov(\beta_k, \sigma^2|\mathbf{y}, \mathbf{X}) = 0. \tag{44}$$

These closed-form expressions are used to calculate DIC_L and DIC_M. For comparison, we also calculate the BF of M_k against M_1 when the g-prior is used for both θ_k and θ_1. The BF has a closed-form expression given by

$$\text{BF}\left(M_k, M_1\right) = \frac{(1+g)^{(n-k-1)/2}}{\left(1+g\left(1-R_k^2\right)\right)^{(n-1)/2}}, \tag{45}$$

where $R_k^2 = 1 - \frac{(\mathbf{y}-\mathbf{X}_k\hat{\beta}_k)'(\mathbf{y}-\mathbf{X}_k\hat{\beta}_k)}{(\mathbf{y}-\bar{\mathbf{y}})'(\mathbf{y}-\bar{\mathbf{y}})}$ with $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^n y_i$. For the details about the g-prior and the BF, see Liang et al. (2008).

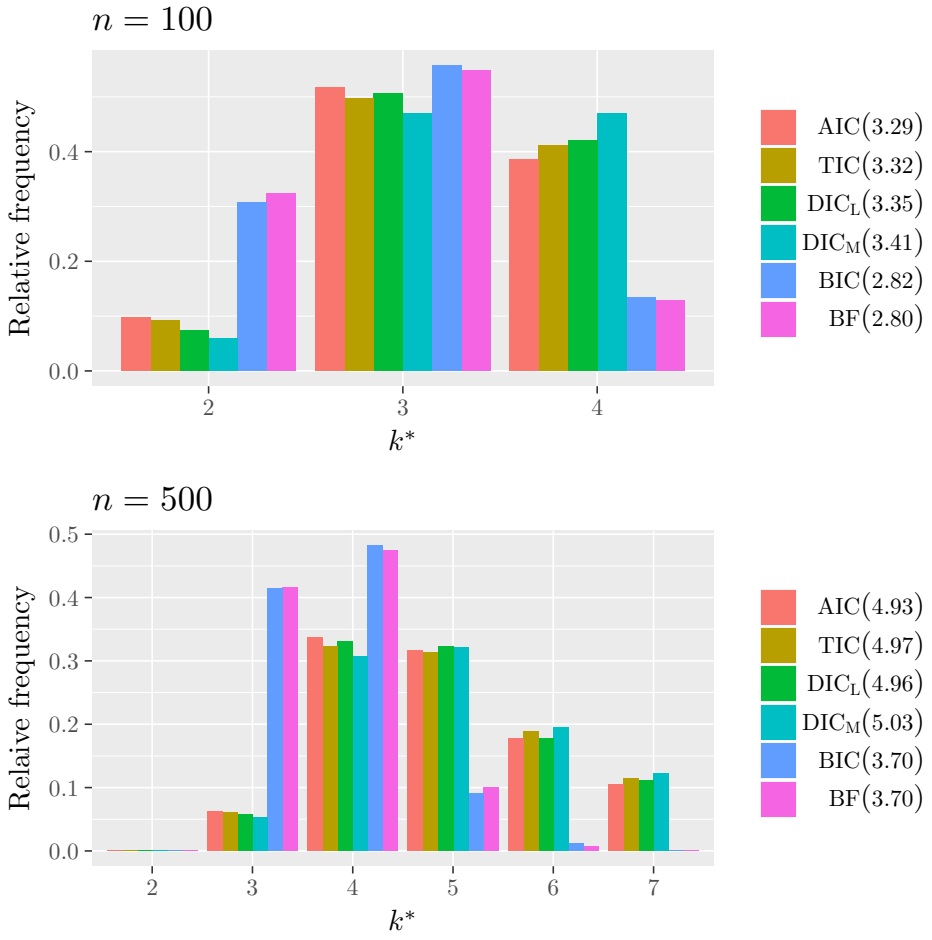
Each of the six criteria is used to select the best model (call it M_{k^*}). Based on M_{k^*}, we then calculate $EKL(k^*)$ where $EKL(k^*)$ is $EKL_{ML}(k^*)$ defined in Equation (5) for AIC, TIC, and BIC and is $EKL_B(k^*)$ defined in Equation (7) for DIC_L, DIC_M, and the BF. In general $EKL(k^*)$ does not have a closed-form expression and a numerical method is needed. To compute $EKL(k^*)$, we first simulate $1,000$ replications of $\mathbf{y}$ from M_k, denoted by $\mathbf{y}^l$ for $l = 1, 2, \ldots, 1,000$. Then, for each $\mathbf{y}^l$, we simulate $1,000$ replications of $\mathbf{y}^l$ from M_k, denoted by $\mathbf{y}_{rep}^m$ for $m = 1, 2, \ldots, 1,000$. These simulations are possible here because we know what the true DGP is. Then we calculate $EKL(k^*)$ by

$$\begin{aligned}
\widehat{EKL}_{ML}(k^*) &= \frac{1}{1000}\sum_{l=1}^{1000}\frac{1}{1000}\sum_{m=1}^{1000} D\left(\mathbf{y}_{rep}^m|\widehat{\theta}_{k^*}\left(\mathbf{y}^l\right), M_{k^*}\right), \text{ for AIC, TIC, BIC;} \\
\widehat{EKL}_B(k^*) &= \frac{1}{1000}\sum_{l=1}^{1000}\frac{1}{1000}\sum_{m=1}^{1000} D\left(\mathbf{y}_{rep}^m|\overline{\theta}_{k^*}\left(\mathbf{y}^l\right), M_{k^*}\right), \text{ for } \text{DIC}_L, \text{ DIC}_M, \text{ BF.}
\end{aligned}$$

The relative frequencies of the selected models by each of six criteria (namely AIC, TIC, DIC_L, DIC_M, BF, and BIC) are reported in Figure 1. Also reported in Figure 1 are the average values of k^*, all across $1,000$ replications. Several interesting results can be found in Figure 1. The models selected by the BF and BIC tend to be more parsimonious than those selected by AIC, TIC, DIC_L and DIC_M. This result is not surprising as BIC has a larger penalty term than AIC. Second, the average k^*s selected by the BF and BIC are very similar to each other, suggested that they tend to select the same model, especially when $n = 500$. Similarly, the average k^*s selected by AIC and DIC_L are very similar, suggested that they tend to select the same model. Also, the average k^*s selected by TIC and DIC_M are very similar, suggested that they tend to select the same model. Third, as the sample size increases, the average k^*s selected by all criteria, including BIC and the BF, tend to increase. This is not surprising as the true DGP is not a candidate model.

Table 1 reports the average values of $(EKL(k^*) - 1 - \ln(2\pi))$, scaled by 1,000, where $EKL(k^*)$ is $EKL_{ML}(k^*)$ for AIC, TIC, and BIC and $EKL_B(k^*)$ for DIC_L, DIC_M, and the BF, all across $1,000$ replications. We report $(EKL(k^*)-1-\ln(2\pi))\times 10^3$ instead of $EKL(k^*)$ to better highlight differences in the expected KL divergence under different criteria. The most important result from Table 1 is that DIC_M leads to a much smaller value of the expected

Figure 1: The figure plots relative frequencies of the polynomial orders selected by different criteria. The numbers in parentheses are the average values of k^*s.



KL divergence than the BF when $n = 100$ and 500. Even though DIC_L is not asymptotically justified in this case due to the omission of the true DGP in the set of candidate models, DIC_L leads to a small value of the expected KL divergence than the BF. Interestingly and not surprisingly, TIC leads to a small value of the expected KL divergence than BIC. Results obtained from this Monte Carlo study indicate that if one's objective is to choose a model that leads to a smaller value for the KL divergence between the DGP and $p\left(\mathbf{y}_{rep}|\bar{\theta}(\mathbf{y})\right)$, it is better to use DIC_M than the BF. Similarly, if one's objective is to choose a model that leads to a smaller value for the KL divergence between the DGP and $p\left(\mathbf{y}_{rep}|\widehat{\theta}(\mathbf{y})\right)$, it is better to use TIC than BIC.

Table 1: The average value of $(EKL(k^*) - 1 - \ln(2\pi))$, scaled by 1,000, across 1,000 replications, under different criteria

Criteria	AIC	TIC	DIC_L	DIC_M	BIC	BF
$n = 100$	67.80293	67.47593	61.38793	60.36793	79.35193	76.49893
$n = 500$	15.68993	15.67293	15.27493	15.20993	20.13193	19.98293

5 Applications

We now illustrate the proposed method in two applications. The first example is asset pricing models under the Student t distribution. The likelihood functions of these models not only have analytical form, but also can be rewritten in a latent variable form. We choose this example to compare the two alternative formulations of the same model, paying particular attention to the impact the two equivalent formulations on DIC, DIC_L, DIC_M. In the second example $p\left(\mathbf{y}|\bar{\theta}\right)$ is not available in closed-form. Given that DIC_1 is too difficult to compute, we calculate DIC_L and DIC_M by particle filters proposed in Section 3.4.3.

5.1 Factor asset pricing models

Factor asset pricing models are important in modern finance. There models generally assume that the return distribution is normal. Unfortunately, there has been overwhelming empirical evidence against normality for asset returns, which have led researchers to investigate asset pricing models with heavy-tailed distributions. Zhou (1993) suggested using the multivariate t distribution to replace the multivariate normal distribution. Moreover, based on the efficient market theory, the asset excess premium should not be statistically different from zero. At last, the multivariate t distribution can be rewritten as scale-mixture framework to become a latent variable model. Hence, we consider the following six asset pricing models:

$$
\begin{aligned}
&\text{Model 1:} R_t = \beta \mathbf{F}_t + \epsilon_t, \epsilon_t \sim N[\mathbf{0}, \mathbf{\Sigma}], \\
&\text{Model 2:} R_t = \alpha + \beta \mathbf{F}_t + \epsilon_t, \epsilon_t \sim N[\mathbf{0}, \mathbf{\Sigma}], \\
&\text{Model 3:} R_t = \beta \mathbf{F}_t + \epsilon_t, \epsilon_t \sim t[\mathbf{0}, \mathbf{\Sigma}, \nu], \\
&\text{Model 4:} R_t = \beta \mathbf{F}_t + \epsilon_t, \epsilon_t \sim N(\mathbf{0}, \mathbf{\Sigma}/\omega_t), \omega_t \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \\
&\text{Model 5:} R_t = \alpha + \beta \mathbf{F}_t + \epsilon_t, \epsilon_t \sim t[\mathbf{0}, \mathbf{\Sigma}, \nu], \\
&\text{Model 6:} R_t = \alpha + \beta \mathbf{F}_t + \epsilon_t, \epsilon_t \sim N(\mathbf{0}, \mathbf{\Sigma}/\omega_t), \omega_t \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right),
\end{aligned}
$$

where R_t is the excess return of portfolio at period t with $N \times 1$ dimension, $\mathbf{F}_t$ a $K \times 1$ vector of factor portfolio excess returns, α a $N \times 1$ vector of intercepts, β a $N \times K$ vector of scaled covariances, ϵ_t the random error, $t = 1, 2, \cdots, n$. For convenience, we restrict $\mathbf{\Sigma}$ to be a diagonal matrix and ν to be a known constant as $\nu = 3$. It is noted that Model 4 is the scale-mixture distributional representation of Model 3, and Model 5 is the scale mixture distributional representation of Model 6.

Monthly returns of 25 portfolios, constructed at the end of each June, are the intersections of 5 portfolios formed on size (market equity, ME) and 5 portfolios formed on the ratio of book equity to market equity (BE/ME). The Fama/French's three factors, market excess return, SMB (Small Minus Big), HML (High Minus Low) are used as the explanatory factors (Fama and French, 1993). The sample period is from July 1926 to November 2017, so that $N = 25$, $n = 1097$. The data are freely available from the data library of Kenneth French.[5]

Bayesian inference for factor asset pricing models has attracted a considerable amount of attention in the empirical asset pricing literature. Avramov and Zhou (2010) provided an excellent review of the literature on Bayesian portfolio analysis. To obtain MCMC output, we need specify the prior distributions for parameters. Here, to represent the prior ignorance, we assign some vague conjugate prior distributions,

$$\alpha_i \sim N[0, 100], \beta_{ij} \sim N[0, 100], \Sigma_{ii}^{-1} \sim \Gamma[0.01, 0.01].$$

Here, we draw 100,000 random observations from the posterior distributions in each model where the first 40,000 is used as the burn-in sample, and the next 60,000 iterations is collected with every 3^{rd} observation as effective observations. Hence, these are 20,000 effective observations.

To compare these models, based on 20,000 effective observations, we calculate DIC_1, $P_{D,1}$, DIC_L, P_L, DIC_M, P_M, for all candidate models, and DIC_7 and $P_{D,7}$ for Model 4 and Model 6 as there are latent variables in these two models. The results are reported in Table 2. Several interesting findings emerge from Table 2. First, DIC_1 in Model 3 is very different from DIC_7 in Model 4, although these two models are the same. The reason for the difference is that in Model 3 there is no latent variable, whereas in Model 4 the scale-mixture representation of the Student t distribution introduces latent variables, $\{\omega_t\}$. Due to the difference, the common practice of DIC for Model 3 is DIC_1 and for Model 4 is DIC_7. The sharp difference between the two DIC values for the identical model is clearly unsatisfactory. For the same reason, DIC_1 in Model 5 is very different from DIC_7 in Model 6. Second, the asymptotic results developed in Li et al. (2017) and in Theorem 3.1 above suggest that $P_{D,1}$ and P_L should be close to the actual number of the parameters, P, if the prior distribution is dominated by the likelihood function. The results are confirmed by Table 2. Not surprisingly, $P_{D,1}$ is almost identical to P_L and DIC_1 and DIC_L are almost the same for each candidate model. Finally, DIC, DIC_L, DIC_M, all pick Model 6 (and Model 5) as the best model.

5.2 Stochastic volatility models

Stochastic volatility (SV) models have been found very useful for pricing derivative securities. In the discrete-time log-normal SV models, the lo-volatility is the state variable which is often

[5] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Table 2: Model selection results for Fama-French three factor models

Model	M_1	M_2	M_3	M_4	M_5	M_6
P	100	125	100	100	125	125
$P_{D,1}$	100	125	100	100	125	125
DIC_1	-132196	-132762	-143510	-143510	-144635	-144635
$P_{D,7}$	NA	NA	NA	1090	NA	1115
DIC_7	NA	NA	NA	-145159	NA	-146339
P_L	100	125	100	100	126	126
DIC_L	-132196	-132762	-143509	-143509	-144634	-144634
P_M	997	1015	291	291	403	403
DIC_M	-130402	-130982	-143128	-143128	-144079	-144079

assumed to follow an AR(1) model. The basic log-normal SV model is of the form:

$$\begin{aligned} y_t &= \exp(h_t/2)u_t,\ u_t \sim N(0,1), \\ h_t &= \mu + \phi(h_{t-1} - \mu) + \tau v_t,\ v_t \sim N(0,1), \end{aligned}$$

where $t = 1, 2, \cdots, n$, y_t is the continuously compounded return, h_t the unobserved log-volatility, $h_0 = \mu$, u_t and v_t are independent for all t. In this paper, we denote this model M_1.

To carry out MCMC analysis of M_1, following Meyer and Yu (2000), the prior distributions are specified as follows:

$$\mu \sim N(0, 100), \phi \sim Beta(1,1), \quad 1/\tau^2 \sim \Gamma(0.001, 0.001).$$

An important and well documented empirical feature in many financial time series is the leverage effect. Following Yu (2005), we define the leverage effect SV model as:

$$\begin{aligned} y_t &= \exp(h_t/2)\, u_t,\ u_t \sim N(0,1) \\ h_{t+1} &= \mu + \phi(h_t - \mu) + \tau v_{t+1}, v_{t+1} \sim N(0,1) \end{aligned}$$

with

$$\begin{pmatrix} u_t \\ v_{t+1} \end{pmatrix} \overset{i.i.d}{\sim} N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}$$

and $h_0 = \mu$. In this model, ρ captures the leverage effect if $\rho < 0$. In this case, there is a negative relationship between the expected future volatility and the current return. We denote this model M_2 and specify the prior distribution of ρ as $\rho \sim$Unif$(-1, 1)$.

Our goal here is to compare the two models using DIC_7, DIC_L and DIC_M. In both cases, $p(\mathbf{y}|\theta)$ is not available in closed-form. Since both specifications are nonlinear non-Gaussian state-space models, the Kalman filter is not applicable, making DIC_1 is time-consuming to

compute. To compute DIC_L and DIC_M, we use the particle filters to evaluate the observed-data likelihood and its second derivatives.

The dataset consists of 945 daily mean-corrected returns on Pound/Dollar exchange rates, covering the period between 01/10/81 and 28/06/85. For MCMC, after a burn-in period of 10,000 iterations, we save every 20th value for the next 100,000 iterations to get 5,000 effective draws. The same dataset was used in Kim et al. (1998) and Meyer and Yu (2000). The posterior mean and standard error of parameters in the two competing model are reported in Table 3. Note that the in M_2, the posterior mean of ρ is very close to zero, relative to its posterior standard error.

Table 3: Posterior mean and standard error of parameters in M_1 and M_2

	M_1		M_2	
Parameter	Mean	SE	Mean	SE
μ	-0.6733	0.3282	-0.6485	0.3377
ϕ	0.9733	0.0127	0.9802	0.0138
ρ	NA	NA	-0.0575	0.1570
τ	0.1698	0.0378	0.1661	0.0391

Table 4 reports DIC_7, $P_{D,7}$, DIC_L, P_L, DIC_M, P_M. The following findings can be obtained from Table 3. First and foremost, DIC_L and DIC_L^{BP} suggest the same ranking of the competing models, but DIC_7 is different. In particular, DIC_7 suggests that M_2 is better than M_1. According to DIC_7, M_1 and M_2 perform nearly the same judged by $D\left(\overline{\theta}\right)$. However, M_2 reduces the effective number of parameters by 22.3 over M_1. This reduction of the model complexity is the reason why DIC_7 prefers M_2. This result is surprising as the posterior mean of the leverage effect is nearly zero, as reported in Table 2. On the other hand, DIC_L suggests that M_1 is slightly better than M_2 although the difference is not worth to mention. In DIC_L, P_L is 2.32 in M_1 and 3.24 in M_2. These values are very close to the actual numbers of parameters in the two models. Similar results are found in DIC_M. P_M is 4.44 in M_1 and 5.02 in M_2. Given that M_2 has one extra parameter, this difference is reasonable. Moreover, M_1 and M_2 perform nearly the same judged by $D\left(\overline{\theta}\right)$. These two observations explain why M_1 is slightly better that M_2. This empirical example clearly demonstrates that DIC_L and DIC_M can select more reasonable models than DIC_7. We can compare the computational time. The CPU time for computing DIC_L and DIC_M together is 345 seconds.[6] For DIC_1, the CPU time is 1922 seconds. If one increases the number of effective draws, the CPU time will increase linearly for DIC_1 but remain the same order for DIC_L and DIC_M.

[6]The CPU time is based on Laptop Intel (R) Core (TM) i7-7500H CPU @2.70GHz, implementing MATLAB R2017b.

Table 4: Model selection results for M_1 and M_2

Model	$P_{D,7}$	$D(\bar{\theta})$	DIC_7	P_L	$D(\bar{\theta})$	DIC_L	P_M	$D(\bar{\theta})$	DIC_M
M_1	53.60	1695.40	1802.52	2.32	1837.81	1842.50	4.44	1837.81	1846.69
M_2	31.33	1693.36	1756.21	3.24	1837.78	1844.30	5.02	1837.78	1847.82

6 Conclusion

Although latent variable models can be conveniently estimated in the Bayesian framework via MCMC if the data augmentation technique is used, we argue that the conditional likelihood function should not be used to obtain DIC. This is because, the conditional likelihood invalidate the standard Bayesian large sample theory and the ML asymptotic theory, which are needed to show that DIC is an asymptotically unbiased estimator of the expected KL divergence between the DGP and the predictive distribution. An example is given where DIC provides an asymptotically biased estimator of the expected KL divergence between the DGP and the predictive distribution.

While in principle one can use the standard DIC (i.e. DIC_1), in practice, DIC_1 is very difficult to calculate for many latent variable models because the observed-data likelihood is not available in closed-form. In particular, one has to numerically evaluate the observed-data likelihood at each MCMC iteration. It makes the implementation of DIC_1 practically non-operational for many latent variable models.

We introduce DIC_L for comparing latent variable models. We show that DIC_L can be justified by the standard Bayesian asymptotic theory. In particular, we show that DIC_L is an asymptotically unbiased estimator of the expected KL divergence minus $2C$ when the loss function is based on a plug-in predictive distribution. We then develop a simple and general approach to computing DIC_L for latent variable models. Since the latent variables are not treated as parameters in defining DIC_L, DIC_L is robust to nonlinear transformations of the latent variables.

The justification of DIC_1 and DIC_L requires the candidate model is a good approximation to the true DGP. We develop DIC_M to compare misspecified models. DIC_M can be regarded as the Bayesian version of TIC. Under a set of regularity conditions, we show that DIC_M is an asymptotically unbiased estimator of the expected KL divergence minus $2C$ when the loss function is based on a plug-in predictive distribution. The advantages of DIC_L and DIC_M are illustrated using two popular models. Empirical examples demonstrates that DIC_L and DIC_M can select more reasonable models than DIC_7, a widely-used Bayesian model selection criterion to compare latent variables. The detail of the implementation of DIC_L and DIC_M can be found in Li et al. (2019) where the R code may be downloaded.

Appendix

6.1 Notations

$:=$	definitional equality	$\xrightarrow{p}$	converge in probability
$o(1)$	tend to zero	$\hat{\theta}$	ML estimate
$o_p(1)$	tend to zero in probability	θ_n^p	pseudo true parameter
$\overline{\theta}$	posterior mean	DIC_1	DIC based on $p(\mathbf{y}\|\theta)$
DIC_7	DIC based on $p(\mathbf{y}\|\theta,\mathbf{z})$	DIC_L	DIC for latent variable models
DIC_M	DIC for misspecified models		

Proof of Lemma 3.1

We can decompose $\frac{1}{n}\sum_{t=1}^{n}\left[l_t\left(\theta\right)-l_t\left(\theta_n^p\right)\right]$ as

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^{n}\left[l_t\left(\theta\right)-l_t\left(\theta_n^p\right)\right]\\ =\;&\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(\theta\right)-E\left[l_t\left(\theta\right)\right]\right)+\frac{1}{n}\sum_{t=1}^{n}\left(E\left[l_t\left(\theta\right)\right]-E\left[l_t\left(\theta_n^p\right)\right]\right)+\frac{1}{n}\sum_{t=1}^{n}\left(E\left[l_t\left(\theta_n^p\right)\right]-l_t\left(\theta_n^p\right)\right).\end{aligned}$$

From (18), we know that for any $\varepsilon>0$, there exists $\delta_1(\varepsilon)>0$ and $N(\varepsilon)>0$, for all $n>N\left(\varepsilon\right)$,

$$\frac{1}{n}\sum_{t=1}^{n}\left\{E\left[l_t\left(\theta\right)\right]-E\left[l_t\left(\theta_n^p\right)\right]\right\}<-\delta_1(\varepsilon),$$

if $\theta\in\mathbf{\Theta}\backslash N\left(\theta_n^p,\varepsilon\right)$. Thus, for any $\varepsilon>0$, if $\theta\in\mathbf{\Theta}\backslash N\left(\theta_n^p,\varepsilon\right)$, for all $n>N\left(\varepsilon\right)$,

$$\frac{1}{n}\sum_{t=1}^{n}\left[l_t\left(\theta\right)-l_t\left(\theta_n^p\right)\right]<\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(\theta\right)-E\left[l_t\left(\theta\right)\right]\right)-\delta_1\left(\varepsilon\right)+\frac{1}{n}\sum_{t=1}^{n}\left(E\left[l_t\left(\theta_n^p\right)\right]-l_t\left(\theta_n^p\right)\right),$$

and

$$\begin{aligned}&\sup_{\mathbf{\Theta}\backslash N\left(\theta_n^p,\varepsilon\right)}\frac{1}{n}\sum_{t=1}^{n}\left[l_t\left(\theta\right)-l_t\left(\theta_n^p\right)\right]\\ \leq\;&\sup_{\mathbf{\Theta}\backslash N\left(\theta_n^p,\varepsilon\right)}\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(\theta\right)-E\left[l_t\left(\theta\right)\right]\right)-\delta_1\left(\varepsilon\right)+\frac{1}{n}\sum_{t=1}^{n}\left(E\left[l_t\left(\theta_n^p\right)\right]-l_t\left(\theta_n^p\right)\right)\\ \leq\;&\sup_{\mathbf{\Theta}\backslash N\left(\theta_n^p,\varepsilon\right)}\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(\theta\right)-E\left[l_t\left(\theta\right)\right]\right)\right|-\delta_1\left(\varepsilon\right)+\left|\frac{1}{n}\sum_{t=1}^{n}\left(E\left[l_t\left(\theta_n^p\right)\right]-l_t\left(\theta_n^p\right)\right)\right|\\ \leq\;&2\sup_{\theta\in\mathbf{\Theta}}\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(\theta\right)-E\left[l_t\left(\theta\right)\right]\right)\right|-\delta_1\left(\varepsilon\right).\end{aligned}\tag{46}$$

Under Assumptions 1-6, the uniform convergence condition is satisfied, that is,

$$P\left(\sup_{\theta\in\mathbf{\Theta}}\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(\theta\right)-E\left[l_t\left(\theta\right)\right]\right)\right|<\varepsilon\right)\to1,\tag{47}$$

From the uniform convergence, if we choose δ_2 such that $0 < \delta_2 < \delta_1(\varepsilon)/2$, we have

$$P\left[\sup_{\theta\in\mathbf{\Theta}}\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t(\theta) - E\left[l_t(\theta)\right]\right)\right| < \delta_2\right] \to 1.$$

Hence,

$$P\left[2\sup_{\theta\in\mathbf{\Theta}}\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t(\theta) - E\left[l_t(\theta)\right]\right)\right| - \delta_1(\varepsilon) < 2\delta_2 - \delta_1(\varepsilon)\right] \to 1.$$

From (46), we have

$$\begin{aligned} & P\left[2\sup_{\theta\in\mathbf{\Theta}}\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t(\theta) - E\left[l_t(\theta)\right]\right)\right| - \delta_1(\varepsilon) < 2\delta_2 - \delta_1(\varepsilon)\right] \\ \leqslant\ & P\left[\sup_{\mathbf{\Theta}\setminus N\left(\theta_n^p,\varepsilon\right)}\frac{1}{n}\left[\sum_{t=1}^{n}l_t(\theta) - \sum_{t=1}^{n}l_t\left(\theta_n^p\right)\right] < 2\delta_2 - \delta_1(\varepsilon)\right]. \end{aligned}$$

Letting $K_1(\varepsilon) = -\left(2\delta_2 - \delta_1(\varepsilon)\right) > 0$, we have, for any ε,

$$\lim_{n\to\infty} P\left[\sup_{\mathbf{\Theta}\setminus N\left(\theta_n^p,\varepsilon\right)}\frac{1}{n}\left[\sum_{t=1}^{n}l_t(\theta) - \sum_{t=1}^{n}l_t\left(\theta_n^p\right)\right] < -K_1(\varepsilon)\right] = 1,$$

which proves the consistency condition given by (19). The proof of the other two concentration conditions (20) can be done similarly and hence omitted.

6.2 Proof of Lemma 3.2

In this subsection, for any function $f(\theta)$, let $f^{(j)}(\theta)$ be the jth order derivative of $f(\theta)$ for $j = 1, 2, 3, 4, 5$. Furthermore, let $\hat{f}$ be the value of function f evaluated at $\hat{\theta}$, that is, $\hat{f} := f\left(\hat{\theta}\right)$ and for convenience of exposition, we write $\frac{\partial^d}{\partial\theta_{j_1}\partial\theta_{j_2}\cdots\partial\theta_{j_d}}f(\theta)$ as $f_{j_1\cdots j_d}$ and let $\hat{f}_{j_1\cdots j_d} := f_{j_1\cdots j_d}\left(\hat{\theta}\right)$. For the definition of high order derivatives, we follow Magnus and Neudecker (1999), except that the first-order derivative of a scalar function in our setting is a column vector. Then the Hessian matrix at θ is denoted by $h_n^{(2)}(\theta)$ which is briefly written as $h^{(2)}$ and its (i,j)-component is written as h_{ij} while the components of its inverse is written as σ_{ij}. Let μ^4_{ijkq}, μ^6_{ijkqrs}, $\mu^8_{ijkqrstw}$, $\mu^{10}_{ijkqrstwv\beta}$, $\mu^{12}_{ijkqrstwv\beta\tau\phi}$ be the fourth, sixth, eighth, tenth, and twelfth central moments of a multivariate Normal distribution whose covariance matrix is $\hat{h}^{(-2)} := \left(h^{(2)}(\theta)\right)^{-1}|_{\theta=\hat{\theta}}$.

We say the pair $(\{h_n\}, b)$ satisfies the analytical assumptions for the stochastic Laplace method on $\wp_\theta$, if the following assumptions are met. There exists positive numbers ε, M and η such that (i) with probability approach one (w.p.a.1), for all $\theta \in B_\varepsilon\left(\hat{\theta}\right)$ and all $1 \leq j_1, \cdots, j_d \leq P$ with $0 \leq d \leq 8$, $\|h_n(\theta)\| < M$ and $\|h_{j_1\cdots j_d}(\theta)\| < M$; (ii) w.p.a.1, $\hat{h}^{(2)}$ is positive definite and $\det\left(\hat{h}^{(2)}\right) > \eta$; (iii) For all $\varepsilon > 0$, there exists $K_1(\varepsilon) > 0$,

$\sup_{\mathbf{\Theta}\setminus B\left(\theta_{n,\varepsilon}^{p}\right)} \frac{1}{n}\left[-h_n\left(\theta\right)-\left(-h_n\left(\theta_n^p\right)\right)\right] < -K_1\left(\varepsilon\right)$, w.p.a.1; (iv) w.p.a.1, for all $\theta \in B_\varepsilon\left(\hat{\theta}\right)$ and all $1 \leq j_1, \cdots, j_d \leq P$, with $0 \leq d \leq 6$, $\left\|b\left(\theta\right)\right\| < M$ and $\left\|b_{j_1 \cdots j_d}\left(\theta\right)\right\| < M$.

Note that our assumptions are different from those in Section 3 of Kass et al. (1990) in two aspects. First, we require $h_n(\theta)$ be eight-times continuously differentiable and $b(\theta)$ be six-times continuously differentiable. Second, for conditions (ii) and (iii), instead of almost sure boundedness and almost sure convergence, we assume they hold w.p.a.1. We do so because we are interested in convergence in probability only. To prove Lemma 3.2, we first review a result of Li et al. (2017).

Lemma 6.1 *For some real-valued function $g(\theta)$, if both $\left(\left\{h_n(\theta)\right\}, g(\theta)b_D(\theta)\right)$ and $\left(\left\{h_n(\theta)\right\}, b_D(\theta)\right)$ satisfy the analytical assumptions for the stochastic Laplace method on $\wp_\theta$, then*

$$\frac{\int g\left(\theta\right) b_D\left(\theta\right) \exp\left(-nh_n\left(\theta\right)\right) d\theta}{\int b_D\left(\theta\right) \exp\left(-nh_n\left(\theta\right)\right) d\theta} = \hat{g} + \frac{1}{n}B_1 + \frac{1}{n^2}\left(B_2 - B_3\right) + O_p\left(\frac{1}{n^3}\right),$$

where

$$B_1 = \frac{1}{2}\sum_{ij}\hat{\sigma}_{ij}\hat{g}_{ij} + \frac{\sum_{ij}\hat{\sigma}_{ij}\hat{b}_{D,j}\hat{g}_i}{\hat{b}_D} - \frac{1}{6}\sum_{ijkq}\hat{h}_{ijk}\mu^4_{ijkq}\hat{g}_q,$$

$$\begin{aligned}
B_2 \;=\; & -\frac{1}{120}\sum_{ijkqrs}\hat{h}_{ijkqr}\mu^6_{ijkqrs}\hat{g}_s + \frac{1}{144}\sum_{ijkqrstw}\hat{h}_{ijk}\hat{h}_{qrst}\mu^8_{ijkqrstw}\hat{g}_w \\
& -\frac{1}{1296}\sum_{ijkqrstwv\beta}\hat{h}_{ijk}\hat{h}_{qrs}\hat{h}_{twv}\mu^{10}_{ijkqrstwv\beta}\hat{g}_\beta - \frac{1}{24}\frac{\sum_{ijkqrs}\hat{h}_{ijkq}\mu^6_{ijkqrs}\hat{b}_{D,s}\hat{g}_r}{\hat{b}_D} \\
& +\frac{1}{72}\frac{\sum_{ijkqrstw}\hat{h}_{ijk}\hat{h}_{qrs}\mu^8_{ijkqrstw}\hat{b}_{D,w}\hat{g}_t}{\hat{b}_D} - \frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\hat{h}_{ijk}\mu^6_{ijk\zeta\eta\xi}\hat{b}_{D,\eta\xi}\hat{g}_\zeta}{\hat{b}_D} \\
& +\frac{1}{6}\frac{\sum_{\zeta\eta\xi\omega}\mu^4_{\zeta\eta\xi\omega}\hat{b}_{D,\eta\xi\omega}\hat{g}_\zeta}{\hat{b}_D} - \frac{1}{48}\sum_{ijkqrs}\hat{h}_{ijkq}\mu^6_{ijkqrs}\hat{g}_{rs} \\
& +\frac{1}{144}\sum_{ijkqrstw}\hat{h}_{ijk}\hat{h}_{qrs}\mu^8_{ijkqrstw}\hat{g}_{tw} - \frac{1}{36}\sum_{ijk\zeta\eta\xi}\hat{h}_{ijk}\mu^6_{ijk\zeta\eta\xi}\hat{g}_{\zeta\eta\xi} \\
& +\frac{1}{24}\sum_{\zeta\eta\xi\omega}\mu^4_{\zeta\eta\xi\omega}\hat{g}_{\zeta\eta\xi\omega} - \frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\hat{h}_{ijk}\mu^6_{ijk\zeta\eta\xi}\hat{g}_{\zeta\eta}\hat{b}_{D,\xi}}{\hat{b}_D} \\
& +\frac{1}{6}\frac{\sum_{\zeta\eta\xi\omega}\mu^4_{\zeta\eta\xi\omega}\hat{g}_{\zeta\eta\xi}\hat{b}_{D,\omega}}{\hat{b}_D} + \frac{1}{4}\frac{\sum_{\zeta\eta\xi\omega}\mu^4_{\zeta\eta\xi\omega}\hat{g}_{\zeta\eta}\hat{b}_{D,\xi\omega}}{\hat{b}_D},
\end{aligned}$$

$$B_3 = B_4 \times B_1,$$

$$B_4 = \frac{1}{2}\sum_{ij}\hat{\sigma}_{ij}\frac{\hat{b}_{D,ij}}{\hat{b}_D} - \frac{1}{6}\sum_{ijkq}\hat{h}_{ijk}\mu^4_{ijkq}\frac{\hat{b}_{D,q}}{\hat{b}_D} + \frac{1}{72}\sum_{ijkqrs}\hat{h}_{ijk}\hat{h}_{qrs}\mu^6_{ijkqrs} - \frac{1}{24}\sum_{ijkq}\hat{h}_{ijkq}\mu^4_{ijkq}.$$

Lemma 6.2 (The Generalized Isserlis Theorem) *If $A = \{\alpha_1, \ldots, \alpha_{2N}\}$ is a set of integers such that $1 \leq \alpha_i \leq P$, for each $i \in [1, 2N]$ and $X \in R^P$ is a zero mean multivariate normal random vector then*

$$EX_A = \underset{A}{\Sigma\Pi} E\left(X_i X_j\right), \tag{48}$$

where $X_A = \prod_{\alpha_i \in A} X_{\alpha_i}$ and the notation $\Sigma\Pi$ means summing over all distinct ways of partitioning $X_{\alpha_1}, \ldots, X_{\alpha_{2N}}$ into pairs (X_i, X_j) and each summand is the product of the N pairs. This yields $(2N)!/\left(2^N N!\right) = (2N-1)!!$ terms in the sum where $(2N-1)!!$ is the double factorial such that $(2N-1)!! = (2N-1)(2N-3)\ldots 1$.

The Isserlis theorem, first obtained by Isserlis (1918), expresses the higher-order moments of a zero-mean Gaussian vector in terms of its covariance matrix. The generalized Isserlis theorem is due to Withers (1985) and Vignat (2012). For instance, let $A = \{1, 1, 2, 4\}$, we have

$$EX_A = E\left(X_1^2 X_2 X_4\right) = \underset{A}{\Sigma\Pi} E\left(X_i X_j\right) = E\left(X_1^2\right) E\left(X_2 X_4\right) + 2E\left(X_1 X_2\right) E\left(X_1 X_4\right).$$

Next, we introduce some useful matrix properties about the vectorization operator.

$$\left(B \otimes C\right)\left(D \otimes E\right) = BD \otimes CE \tag{49}$$

for four matrices B, C, D, and E if BD and CE exist.

$$vec\left(BCD\right) = \left(D' \otimes B\right) vec\left(C\right) \tag{50}$$

for three matrices B, C, and D if the product BCD is defined. And the property between the vectorization operator and trace operator

$$\mathbf{tr}\left(A'BCD'\right) = vec\left(A\right)'\left(D \otimes B\right) vec\left(C\right). \tag{51}$$

On the basis of Lemma 6.1, 6.2, (49), (50) and (51) in the following, we prove the Lemma 3.2.

Proof. First, we define a function $\mathbf{g}\left(\theta\right) = \theta$, and each element of $\mathbf{g}\left(\theta\right)$ is given as $g_z\left(\theta\right) = \theta_z$, $z = 1, \ldots, P$. Denote $\mathbf{g}^{(1)}$, a $P \times P$ matrix, is the first-order derivative of $\mathbf{g}$ evaluated at θ and $\mathbf{g}_{\cdot z}^{(1)}$ is the z_{th} column of $\mathbf{g}^{(1)}$. Note that since $\mathbf{g}\left(\theta\right) = \theta$, $\mathbf{g}^{(1)} = \mathbf{I}_P$ which is the $P \times P$ identity matrix.

For $z = 1, \ldots, P$, $g_z\left(\theta\right)$ is a real-valued function. Hence, using Lemma 6.1, we can get that for each z

$$\frac{\int g_z(\theta) b_D\left(\theta\right) \exp\left(-n h_n\left(\theta\right)\right) d\theta}{\int b_D\left(\theta\right) \exp\left(-n h_n\left(\theta\right)\right) d\theta} = g_z(\theta_n) + \frac{1}{n} B_{1,z}^1 + \frac{1}{n^2}\left(B_{2,z}^1 - B_{3,z}^1\right) + O_p\left(\frac{1}{n^3}\right),$$

Then, in the matrix form, we get

$$\frac{\int \mathbf{g}(\theta) b_D(\theta) \exp(-nh_n(\theta))\, d\theta}{\int b_D(\theta) \exp(-nh_n(\theta))\, d\theta} = \mathbf{g}(\hat{\theta}) + \frac{1}{n} B_1^1 + \frac{1}{n^2}\left(B_2^1 - B_3^1\right) + O_p\left(\frac{1}{n^3}\right).$$

For each z, note that $g_{z,ij} = \frac{\partial g_z^2(\theta}{\partial\theta\partial\theta'}|_{ij} = \mathbf{0}_{ij}$. Following Lemma 6.1, we have

$$B_{1,z}^1 = 0 + \sum_{ij} \hat{g}_{z,i}\hat{\sigma}_{ij}\frac{\hat{b}_{D,j}}{\hat{b}_D} - \frac{1}{6}\sum_{ijkq}\hat{h}_{ijk}\mu_{ijkq}^4\hat{g}_{z,q}.$$

Thus, in the matrix form, we have

$$\begin{aligned} B_1^1 &= \sum_{ij}\hat{\mathbf{g}}_{\cdot i}^{(1)}\hat{\sigma}_{ij}\frac{\hat{b}_{D,j}}{\hat{b}_D} - \frac{1}{2}\sum_{ijkq}\hat{\mathbf{g}}_{\cdot q}^{(1)}\hat{h}_{ijk}\hat{\sigma}_{ij}\hat{\sigma}_{kq} = \sum_{ij}\hat{\mathbf{g}}_{\cdot i}^{(1)}\hat{\sigma}_{ij}\frac{\hat{b}_{D,j}}{\hat{b}_D} - \frac{1}{2}\sum_{ijkq}\hat{\mathbf{g}}_{\cdot q}^{(1)}\hat{\sigma}_{qk}\hat{h}_{ijk}\hat{\sigma}_{ij} \\ &= \hat{\mathbf{g}}^{(1)}\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D} - \frac{1}{2}\hat{\mathbf{g}}^{(1)}\hat{h}^{(-2)}\hat{h}^{(3)\prime} vec\left(\hat{h}^{(-2)}\right), \end{aligned} \tag{52}$$

Hence, we get

$$B_1^1 = \hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D} - \frac{1}{2}\hat{h}^{(-2)}\hat{h}^{(3)\prime} vec\left(\hat{h}^{(-2)}\right). \tag{53}$$

Furthermore, for each z

$$\begin{aligned} B_{2,z}^1 &= -\frac{1}{120}\sum_{ijkqrs}\hat{h}_{ijkqr}\mu_{ijkqrs}^6\hat{g}_{z,s} + \frac{1}{144}\sum_{ijkqrstw}\hat{h}_{ijk}\hat{h}_{qrst}\mu_{ijkqrstw}^8\hat{g}_{z,w} \\ &\quad -\frac{1}{1296}\sum_{ijkqrstwv\beta}\hat{h}_{ijk}\hat{h}_{qrs}\hat{h}_{twv}\mu_{ijkqrstwv\beta}^{10}\hat{g}_{z,\beta} - \frac{1}{24}\frac{\sum_{ijkqrs}\hat{h}_{ijkq}\mu_{ijkqrs}^6\hat{b}_{D,s}\hat{g}_{z,r}}{\hat{b}_D} \\ &\quad +\frac{1}{72}\frac{\sum_{ijkqrstw}\hat{h}_{ijk}\hat{h}_{qrs}\mu_{ijkqrstw}^8\hat{b}_{D,w}\hat{g}_{z,t}}{\hat{b}_D} - \frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\hat{h}_{ijk}\mu_{ijk\zeta\eta\xi}^6\hat{b}_{D,\eta\xi}\hat{g}_{z,\zeta}}{\hat{b}_D} \\ &\quad +\frac{1}{6}\frac{\sum_{\zeta\eta\xi\omega}\mu_{\zeta\eta\xi\omega}^4\hat{b}_{D,\eta\xi\omega}\hat{g}_{z,\zeta}}{\hat{b}_D}. \end{aligned}$$

Thus, in the matrix form, we have

$$\begin{aligned} B_2^1 &= -\frac{1}{120}\sum_{ijkqrs}\hat{g}_{\cdot s}\hat{h}_{ijkqr}\mu_{ijkqrs}^6 + \frac{1}{144}\sum_{ijkqrstw}\hat{g}_{\cdot w}\hat{h}_{ijk}\hat{h}_{qrst}\mu_{ijkqrstw}^8 \\ &\quad -\frac{1}{1296}\sum_{ijkqrstwv\beta}\hat{g}_{\cdot\beta}\hat{h}_{ijk}\hat{h}_{qrs}\hat{h}_{twv}\mu_{ijkqrstwv\beta}^{10} - \frac{1}{24}\frac{\sum_{ijkqrs}\hat{g}_{\cdot r}\hat{h}_{ijkq}\mu_{ijkqrs}^6\hat{b}_{D,s}}{\hat{b}_D} \\ &\quad +\frac{1}{72}\frac{\sum_{ijkqrstw}\hat{g}_{\cdot t}\hat{h}_{ijk}\hat{h}_{qrs}\mu_{ijkqrstw}^8\hat{b}_{D,w}}{\hat{b}_D} - \frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\hat{g}_{\cdot\zeta}\hat{h}_{ijk}\mu_{ijk\zeta\eta\xi}^6\hat{b}_{D,\eta\xi}}{\hat{b}_D} \\ &\quad +\frac{1}{6}\frac{\sum_{\zeta\eta\xi\omega}\hat{g}_{\cdot\zeta}\mu_{\zeta\eta\xi\omega}^4\hat{b}_{D,\eta\xi\omega}}{\hat{b}_D}. \end{aligned} \tag{54}$$

We can write each item on the right-hand side of (54) in the matrix form using (48), that is,

$$-\frac{1}{120}\sum_{ijkqrs}\hat{g}_{\cdot s}\hat{h}_{ijkqr}\mu_{ijkqrs}^6 = -\frac{1}{8}\sum_{ijkqrs}\hat{g}_{\cdot s}\hat{\sigma}_{sr}\hat{h}_{ijkqr}\hat{\sigma}_{ij}\hat{\sigma}_{kq} = -\frac{1}{8}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(5)\prime} vec\left[\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)\right],$$

$$
\begin{aligned}
&\frac{1}{144}\sum_{ijkqrstw}\hat{g}_{\cdot w}\hat{h}_{ijk}\hat{h}_{qrst}\mu^{8}_{ijkqrstw}\\
=\;&\frac{1}{4}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(4)\prime}\left[vec\left(\hat{h}^{(-2)}\right)\otimes\left(vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right)'\right]\\
&+\frac{1}{6}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(4)\prime}\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]vec\left(\hat{h}^{(3)}\right)\\
&+\frac{1}{16}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\mathbf{tr}\left[\left(\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)\right)'\hat{h}^{(4)}\right]\\
&+\frac{1}{4}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\left(\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)'\right)\hat{h}^{(4)}\hat{h}^{(-2)}\right),
\end{aligned}
$$

$$
\begin{aligned}
&-\frac{1}{1296}\sum_{ijkqrstwv\beta}\hat{g}_{\cdot\beta}\hat{h}_{ijk}\hat{h}_{qrs}\hat{h}_{twv}\mu^{10}_{ijkqrstwv\beta}\\
=\;&-\frac{3}{8}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\\
&-\frac{1}{4}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]vec\left(\hat{h}^{(3)\prime}\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\right)\\
&-\frac{1}{16}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\left[vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\right]\\
&-\frac{1}{24}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)vec\left(\hat{h}^{(3)}\right)'\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]vec\left(\hat{h}^{(3)}\right),
\end{aligned}
$$

$$
\begin{aligned}
&-\frac{1}{24}\frac{\sum_{ijkqrs}\hat{g}_{\cdot r}\hat{h}_{ijkq}\mu^{6}_{ijkqrs}\hat{b}_{D,s}}{\hat{b}_D}\\
=\;&-\frac{1}{8}\hat{g}^{(1)}\hat{h}^{(-2)}\frac{\hat{b}^{(1)}_D}{\hat{b}_D}\mathbf{tr}\left[\left[\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)\right]\hat{h}^{(4)\prime}\right]-\frac{1}{2}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(4)\prime}\left[vec\left(\hat{h}^{(-2)}\right)\otimes\left(\hat{h}^{(-2)}\frac{\hat{b}^{(1)}_D}{\hat{b}_D}\right)\right],
\end{aligned}
$$

$$
\begin{aligned}
&\frac{1}{72}\frac{\sum_{ijkqrstw}\hat{g}_{\cdot t}\hat{h}_{ijk}\hat{h}_{qrs}\mu^{8}_{ijkqrstw}\hat{b}_{D,w}}{\hat{b}_D}\\
=\;&\frac{1}{8}\hat{g}^{(1)}\hat{h}^{(-2)}\frac{\hat{b}^{(1)}_D}{\hat{b}_D}vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\\
&+\frac{1}{12}\hat{g}^{(1)}\hat{h}^{(-2)}\frac{\hat{b}^{(1)}_D}{\hat{b}_D}vec\left(\hat{h}^{(3)}\right)'\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]vec\left(\hat{h}^{(3)}\right)\\
&+\frac{1}{2}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\frac{\hat{b}^{(1)}_D}{\hat{b}_D}vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right)'\\
&+\frac{1}{4}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\frac{\hat{b}^{(1)}_D}{\hat{b}_D}'\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)'\\
&+\frac{1}{2}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\left(\hat{h}^{(-2)}\otimes\left(\hat{h}^{(-2)}\frac{\hat{b}^{(1)}_D}{\hat{b}_D}\right)'\right)\hat{h}^{(3)}\hat{h}^{(-2)}\right),
\end{aligned}
$$

$$
\begin{aligned}
& -\frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\hat{g}_{\cdot\zeta}\hat{h}_{ijk}\mu^6_{ijk\zeta\eta\xi}\hat{b}_{D,\eta\xi}}{\hat{b}_D} \\
= \; & -\frac{1}{2}\hat{g}^{(1)}\hat{h}^{(-2)}\frac{\hat{b}^{(2)}_D}{\hat{b}_D}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right) - \frac{1}{2}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(3)'}vec\left(\hat{h}^{(-2)}\frac{\hat{b}^{(2)}_D}{\hat{b}_D}\hat{h}^{(-2)}\right) \\
& -\frac{1}{4}\hat{g}^{(1)}\hat{h}^{(-2)}\hat{h}^{(3)'}vec\left(\hat{h}^{(-2)}\right)\mathbf{tr}\left[\frac{\hat{b}^{(2)}_D}{\hat{b}_D}\hat{h}^{(-2)}\right].
\end{aligned}
$$

$$
\frac{1}{6}\frac{\sum_{\zeta\eta\xi\omega}\hat{g}_{\cdot\zeta}\mu^4_{\zeta\eta\xi\omega}\hat{b}_{D,\eta\xi\omega}}{\hat{b}_D} = \frac{3}{6}\sum_{\zeta\eta\xi\omega}\hat{g}_\zeta\hat{\sigma}_{\zeta\eta}\hat{\sigma}_{\xi\omega}\frac{\hat{b}_{D,\eta\xi\omega}}{\hat{b}_D} = \frac{1}{2}\hat{g}^{(1)}\hat{h}^{(-2)}\frac{\hat{b}^{(3)\prime}_D}{\hat{b}_D}\left[vec\left(\hat{h}^{(-2)}\right)\right].
$$

Hence, we have

$$
\begin{aligned}
B^1_2 \; = \; & -\frac{1}{8}\hat{h}^{(-2)}\hat{h}^{(5)\prime}vec\left[\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)\right] \\
& +\frac{1}{4}\hat{h}^{(-2)}\hat{h}^{(4)\prime}\left[vec\left(\hat{h}^{(-2)}\right)\otimes\left(vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right)'\right] \\
& +\frac{1}{6}\hat{h}^{(-2)}\hat{h}^{(4)\prime}\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]vec\left(\hat{h}^{(3)}\right) \\
& +\frac{1}{16}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\mathbf{tr}\left[\left(\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)\right)'\hat{h}^{(4)}\right] \\
& +\frac{1}{4}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\left(\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)'\right)\hat{h}^{(4)}\hat{h}^{(-2)}\right) \\
& -\frac{3}{8}\hat{h}^{(-2)}\hat{h}^{(3)\prime}\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right) \\
& -\frac{1}{4}\hat{h}^{(-2)}\hat{h}^{(3)\prime}\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]vec\left(\hat{h}^{(3)\prime}\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\right) \\
& -\frac{1}{16}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\left[vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\right] \\
& -\frac{1}{24}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)vec\left(\hat{h}^{(3)}\right)'\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]vec\left(\hat{h}^{(3)}\right) \\
& -\frac{1}{8}\hat{h}^{(-2)}\frac{\hat{b}^{(1)}_D}{\hat{b}_D}\mathbf{tr}\left[\left[\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)\right]\hat{h}^{(4)\prime}\right] \\
& -\frac{1}{2}\hat{h}^{(-2)}\hat{h}^{(4)\prime}\left[vec\left(\hat{h}^{(-2)}\right)\otimes\left(\hat{h}^{(-2)}\frac{\hat{b}^{(1)}_D}{\hat{b}_D}\right)\right] \\
& +\frac{1}{8}\hat{h}^{(-2)}\frac{\hat{b}^{(1)}_D}{\hat{b}_D}vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right) \\
& +\frac{1}{12}\hat{h}^{(-2)}\frac{\hat{b}^{(1)}_D}{\hat{b}_D}vec\left(\hat{h}^{(3)}\right)'\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]vec\left(\hat{h}^{(3)}\right) \\
& +\frac{1}{2}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\frac{\hat{b}^{(1)}_D}{\hat{b}_D}vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right)' \\
& +\frac{1}{4}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\frac{\hat{b}^{(1)\prime}_D}{\hat{b}_D}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)'
\end{aligned}
\tag{55}
$$

$$+\frac{1}{2}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\left(\hat{h}^{(-2)}\otimes\left(\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D}\right)'\right)\hat{h}^{(3)}\hat{h}^{(-2)}\right)$$
$$-\frac{1}{2}\hat{h}^{(-2)}\frac{\hat{b}_D^{(2)}}{\hat{b}_D}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)-\frac{1}{2}\hat{h}^{(-2)}\hat{h}^{(3)'}vec\left(\hat{h}^{(-2)}\frac{\hat{b}_D^{(2)}}{\hat{b}_D}\hat{h}^{(-2)}\right)$$
$$-\frac{1}{4}\hat{h}^{(-2)}\hat{h}^{(3)'}vec\left(\hat{h}^{(-2)}\right)\mathbf{tr}\left[\frac{\hat{b}_D^{(2)}}{\hat{b}_D}\hat{h}^{(-2)}\right]+\frac{1}{2}\hat{h}^{(-2)}\frac{\hat{b}_D^{(3)}{}'}{\hat{b}_D}\left[vec\left(\hat{h}^{(-2)}\right)\right].$$

For B_3^1, following Lemma 6.1, note that, for any element z, $B_{4,z}^1 = B_4^1$ which is a constant and independent of z. We have

$$B_3^1 = B_1^1 \times B_4^1, \tag{56}$$

where

$$B_4^1 = \frac{1}{2}\sum_{ij}\hat{\sigma}_{ij}\frac{\hat{b}_{D,ij}}{\hat{b}_D} - \frac{1}{6}\sum_{ijkq}\hat{h}_{ijk}\mu_{ijkq}^4\frac{\hat{b}_{D,q}}{\hat{b}_D} + \frac{1}{72}\sum_{ijkqrs}\hat{h}_{ijk}\hat{h}_{qrs}\mu_{ijkqrs}^6 - \frac{1}{24}\sum_{ijkq}\hat{h}_{ijkq}\mu_{ijkq}^4. \tag{57}$$

We can write each item on the right-hand side of (57) as

$$\frac{1}{2}\sum_{ij}\hat{\sigma}_{ij}\frac{\hat{b}_{D,ij}}{\hat{b}_D} = \frac{1}{2}\mathbf{tr}\left[\hat{h}^{(-2)}\frac{\hat{b}_D^{(2)}}{\hat{b}_D}\right], \tag{58}$$

$$-\frac{1}{6}\sum_{ijkq}\hat{h}_{ijk}\mu_{ijkq}^4\frac{\hat{b}_{D,q}}{\hat{b}_D} = -\frac{3}{6}\sum_{ijkq}\hat{h}_{ijk}\hat{\sigma}_{ij}\hat{\sigma}_{kq}\frac{\hat{b}_{D,q}}{\hat{b}_D} = -\frac{1}{2}vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(-3)}\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D}, \tag{59}$$

$$\frac{1}{72}\sum_{ijkqrs}\hat{h}_{ijk}\hat{h}_{qrs}\mu_{ijkqrs}^6 \tag{60}$$
$$= \frac{1}{8}vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right) + \frac{1}{12}vec\left(\hat{h}^{(3)}\right)'\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]vec\left(\hat{h}^{(3)}\right),$$

$$-\frac{1}{24}\sum_{ijkq}\hat{h}_{ijkq}\mu_{ijkq}^4 = -\frac{3}{24}\sum_{ijkq}\hat{h}_{ijkq}\hat{\sigma}_{ij}\hat{\sigma}_{kq} = -\frac{1}{8}\mathbf{tr}\left[\left[\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)\right]\hat{h}^{(4)\prime}\right]. \tag{61}$$

From (57), (58), (59), (60), (61), in the matrix form, we have

$$\begin{aligned} B_4^1 &= \frac{1}{2}\mathbf{tr}\left[\hat{h}^{(-2)}\frac{\hat{b}_D^{(2)}}{\hat{b}_D}\right] - \frac{1}{2}vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(-3)}\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D} \\ &\quad + \frac{1}{8}vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right) + \frac{1}{12}vec\left(\hat{h}^{(3)}\right)'\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]vec\left(\hat{h}^{(3)}\right) \\ &\quad - \frac{1}{8}\mathbf{tr}\left[\left[\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)\right]\hat{h}^{(4)\prime}\right]. \end{aligned} \tag{62}$$

From (52), (56) and (56), we have

$$\overline{\theta} = \hat{\theta} + \frac{1}{n}B_1^1 + \frac{1}{n^2}\left(B_2^1 - B_3^1\right) + O_p\left(\frac{1}{n^3}\right) = \hat{\theta} + \frac{1}{n}B_1^1 + \frac{1}{n^2}\left(B_2^1 - B_4^1B_1^1\right) + O_p\left(\frac{1}{n^3}\right).$$

This ends the proof for the first part of the lemma.

In the following, we prove the second part of the lemma. Define a function $\mathbf{f}(\theta) = vec\left(\theta\theta'\right)$ which is a $P^2 \times 1$ vector. Hence, we can get the first and second derivatives of $\mathbf{f}$ with respect to θ as $\mathbf{f}^{(1)}(\theta) = \theta \otimes \mathbf{I}_P + \mathbf{I}_P \otimes \theta$ and $\mathbf{f}^{(2)}(\theta) = (\mathbf{K}_{PP} \otimes \mathbf{I}_P)\left[\mathbf{I}_P \otimes vec(\mathbf{I}_P)\right] + \left[vec(\mathbf{I}_P) \otimes \mathbf{I}_P\right]$ following Magnus and Neudecker (1999), where $\mathbf{K}_{mn}$ is a commutation matrix, which is defined by $\mathbf{K}_{mn}vecA = vecA'$ for a $m \times n$ matrix A. If $m = n$, $\mathbf{K}_{mn}$ is simplified as $\mathbf{K}_m$. By properties of commutation matrix, we have

$$\mathbf{K}_{mn}(Y \otimes x) = x \otimes Y, \tag{63}$$

$$\left(Y \otimes x'\right)\mathbf{K}_{sm} = x' \otimes Y, \tag{64}$$

where Y is a $n \times s$ matrix, x is a $m \times 1$ vector. Furthermore, for any matrix A_1 and A_2, if A_1 is a $n \times s$ dimensional matrix and A_2 is a $m \times t$ dimensional matrix, then,

$$\mathbf{K}_{mn}(A_1 \otimes A_2) = (A_2 \otimes A_1)\mathbf{K}_{ts}. \tag{65}$$

More details about matrix properties, one can refer to Magnus and Neudecker (1999).

Following Lemma 6.1, since each element $f_z(\theta)$ is a real-valued function, we have

$$\frac{\int f_z(\theta) b_D(\theta) \exp\left(-nh_n(\theta)\right) d\theta}{\int b_D(\theta) \exp\left(-nh_n(\theta)\right) d\theta} = f_z\left(\hat{\theta}\right) + \frac{1}{n}B_{1,z}^2 + \frac{1}{n^2}\left(B_{2,z}^2 - B_{3,z}^2\right) + O_p\left(\frac{1}{n^3}\right).$$

Again, we can rewrite it in the matrix form,

$$\frac{\int \mathbf{f}(\theta) b_D(\theta) \exp\left(-nh_n(\theta)\right) d\theta}{\int b_D(\theta) \exp\left(-nh_n(\theta)\right) d\theta} = \mathbf{f}(\theta) + \frac{1}{n}B_1^2 + \frac{1}{n^2}\left(B_2^2 - B_3^2\right) + O_p\left(\frac{1}{n^3}\right).$$

For each z, we have

$$B_{1,z}^2 = \frac{1}{2}\sum_{ij}\hat{\sigma}_{ij}\hat{f}_{z,ij} + \sum_{ij}\hat{f}_{z,i}\hat{\sigma}_{ij}\frac{\hat{b}_{D,j}}{\hat{b}_D} - \frac{1}{6}\sum_{ijkq}\hat{h}_{ijk}\mu_{ijkq}^4\hat{f}_{z,q}.$$

Thus, in the matrix form

$$B_1^2 = \frac{1}{2}\left[\mathbf{I}_{P^2} \otimes vec\left(\hat{h}^{(-2)}\right)'\right] vec\left(\hat{\mathbf{f}}^{(2)\prime}\mathbf{K}_{PP^2}\right) + \sum_{ij}\hat{\mathbf{f}}_{\cdot i}^{(1)}\hat{\sigma}_{ij}\frac{\hat{b}_{D,j}}{\hat{b}_D} - \frac{1}{2}\sum_{ijkq}\hat{\mathbf{f}}_{\cdot q}^{(1)}\hat{h}_{ijk}\hat{\sigma}_{ij}\hat{\sigma}_{kq}. \tag{66}$$

Note that

$$\begin{aligned}
&\left[\mathbf{I}_{P^2} \otimes vec\left(\hat{h}^{(-2)}\right)'\right] vec\left(\hat{\mathbf{f}}^{(2)\prime}\mathbf{K}_{PP^2}\right) \\
= &\left[\mathbf{I}_{P^2} \otimes vec\left(\hat{h}^{(-2)}\right)'\right] vec\left(\left[\mathbf{I}_P \otimes vec(\mathbf{I}_P)'\right](\mathbf{K}_{PP} \otimes \mathbf{I}_P)\mathbf{K}_{PP^2}\right) \\
&+ \left[\mathbf{I}_{P^2} \otimes vec\left(\hat{h}^{(-2)}\right)'\right] vec\left(\left[vec(\mathbf{I}_P)' \otimes \mathbf{I}_P\right]\mathbf{K}_{PP^2}\right)
\end{aligned}$$

where

$$\left[\mathbf{I}_{P^2} \otimes vec\left(\sum_{ij} \hat{\sigma}_{ij} e_i e_j'\right)'\right] vec\left(\left[\mathbf{I}_P \otimes vec\left(\mathbf{I}_P\right)'\right] \mathbf{K}_{PP^2} \left(\mathbf{I}_P \otimes \mathbf{K}_{PP}\right)\right) \tag{67}$$

$$\begin{aligned}
&= \sum_{ij} \hat{\sigma}_{ij} \left[\left(\mathbf{I}_{P^2} \otimes e_j' \otimes e_i'\right) \left(\left(\mathbf{I}_P \otimes \mathbf{K}_{PP}\right) \otimes \left[\mathbf{I}_P \otimes vec\left(\mathbf{I}_P\right)'\right]\right)\right] vec(\mathbf{K}_{PP^2}) \\
&= \sum_{ij} \hat{\sigma}_{ij} \left[\left(\left(\mathbf{I}_{P^2} \otimes e_j'\right) \left(\mathbf{I}_P \otimes \mathbf{K}_{PP}\right)\right) \otimes \left(e_i' \left[\mathbf{I}_P \otimes vec\left(\mathbf{I}_P\right)'\right]\right)\right] vec(\mathbf{K}_{PP^2}) \\
&= \sum_{ij} \hat{\sigma}_{ij} \left[\left[\left(\mathbf{I}_P \otimes \mathbf{I}_P \otimes e_j'\right) \left(\mathbf{I}_P \otimes \mathbf{K}_{PP}\right)\right] \otimes \left(e_i' \otimes vec\left(\mathbf{I}_P\right)'\right)\right] vec(\mathbf{K}_{PP^2}) \\
&= \sum_{ij} \hat{\sigma}_{ij} \left[\left[\mathbf{I}_P \otimes \left(e_j' \otimes \mathbf{I}_P\right)\right] \otimes \left(e_i' \otimes vec\left(\mathbf{I}_P\right)'\right)\right] vec(\mathbf{K}_{PP^2}) \\
&= \sum_{ij} \hat{\sigma}_{ij} vec\left[\left(e_i' \otimes vec\left(\mathbf{I}_P\right)'\right) \mathbf{K}_{PP^2} \left(\mathbf{I}_P \otimes e_j \otimes \mathbf{I}_P\right)\right] \\
&= \sum_{ij} \hat{\sigma}_{ij} vec\left[\left(vec\left(\mathbf{I}_P\right)' \otimes e_i'\right) \left(\mathbf{I}_P \otimes e_j \otimes \mathbf{I}_P\right)\right] = \sum_{ij} \hat{\sigma}_{ij} vec\left[\left(vec\left(\mathbf{I}_P\right)' \left(\mathbf{I}_P \otimes e_j\right)\right) \otimes e_i'\right] \\
&= \sum_{ij} \hat{\sigma}_{ij} vec\left[\left(\left(\mathbf{I}_P \otimes e_j'\right) vec\left(\mathbf{I}_P\right)\right)' \otimes e_i'\right] = \sum_{ij} \hat{\sigma}_{ij} vec\left[e_j \otimes e_i'\right] = vec\left(\hat{h}^{(-2)\prime}\right)
\end{aligned}$$

and

$$\left[\mathbf{I}_{P^2} \otimes vec\left(\hat{h}^{(-2)}\right)'\right] vec\left(\left[vec\left(\mathbf{I}_P\right)' \otimes \mathbf{I}_P\right] \mathbf{K}_{PP^2}\right) \tag{68}$$

$$\begin{aligned}
&= \sum_{ij} \hat{\sigma}_{ij} \left[\left(\mathbf{I}_{P^2} \otimes e_j' \otimes e_i'\right) \left(\mathbf{I}_{P^3} \otimes vec\left(\mathbf{I}_P\right)' \otimes \mathbf{I}_P\right)\right] vec(\mathbf{K}_{PP^2}) \\
&= \sum_{ij} \hat{\sigma}_{ij} \left[\left(\mathbf{I}_{P^2} \otimes e_j'\right) \otimes \left(vec\left(\mathbf{I}_P\right)' \otimes e_i'\right)\right] vec(\mathbf{K}_{PP^2}) \\
&= \sum_{ij} \hat{\sigma}_{ij} vec\left[\left(vec\left(\mathbf{I}_P\right)' \otimes e_i'\right) \mathbf{K}_{PP^2} \left(\mathbf{I}_{P^2} \otimes e_j\right)\right] = \sum_{ij} \hat{\sigma}_{ij} vec\left[\sum_s \left(e_s' \otimes e_s' \otimes e_i'\right) \left(e_j \otimes \mathbf{I}_{P^2}\right)\right] \\
&= \sum_{ij} \hat{\sigma}_{ij} vec\left[\sum_s e_s' e_j \left(e_s' \otimes e_i'\right)\right] = \sum_{ij} \hat{\sigma}_{ij} \sum_s vec\left(e_i e_j' e_s e_s'\right) \\
&= \sum_{ij} \hat{\sigma}_{ij} vec\left(e_i e_j' \sum_s e_s e_s'\right) = \sum_{ij} \hat{\sigma}_{ij} vec\left(e_i e_j'\right) = vec\left(\hat{h}^{(-2)}\right)
\end{aligned}$$

by (50) and (51). Then from (67) and (68), we have

$$\left[\mathbf{I}_{P^2} \otimes vec\left(\hat{h}^{(-2)}\right)'\right] vec\left(\hat{\mathbf{f}}^{(2)\prime} \mathbf{K}_{PP^2}\right) = vec\left(\hat{h}^{(-2)}\right) + vec\left(\hat{h}^{(-2)\prime}\right). \tag{69}$$

Moreover, from (66)

$$B_1^2 = vec\left(\hat{h}^{(-2)}\right) + \hat{\mathbf{f}}^{(1)} \hat{h}^{(-2)} \frac{\hat{b}_D^{(1)}}{\hat{b}_D} - \frac{1}{2} \hat{\mathbf{f}}^{(1)} \hat{h}^{(-2)} \hat{h}^{(-3)\prime} vec\left(\hat{h}^{(-2)}\right)$$

$$= \quad vec\left(\hat{h}^{(-2)}\right) + \hat{\mathbf{f}}^{(1)} B_1^1. \tag{70}$$

And for each z

$$\begin{aligned}
B_{2,z}^2 \quad = \quad & -\frac{1}{120}\sum_{ijkqrs}\hat{h}_{ijkqr}\mu^6_{ijkqrs}\hat{f}_{z,s} + \frac{1}{144}\sum_{ijkqrstw}\hat{h}_{ijk}\hat{h}_{qrst}\mu^8_{ijkqrstw}\hat{f}_{z,w} \\
& -\frac{1}{1296}\sum_{ijkqrstwv\beta}\hat{h}_{ijk}\hat{h}_{qrs}\hat{h}_{twv}\mu^{10}_{ijkqrstwv\beta}\hat{f}_{z,\beta} - \frac{1}{24}\frac{\sum_{ijkqrs}\hat{h}_{ijkq}\mu^6_{ijkqrs}\hat{b}_{D,s}\hat{f}_{z,r}}{\hat{b}_D} \\
& +\frac{1}{72}\frac{\sum_{ijkqrstw}\hat{h}_{ijk}\hat{h}_{qrs}\mu^8_{ijkqrstw}\hat{b}_{D,w}\hat{f}_{z,t}}{\hat{b}_D} - \frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\hat{h}_{ijk}\mu^6_{ijk\zeta\eta\xi}\hat{b}_{D,\eta\xi}\hat{f}_{z,\zeta}}{\hat{b}_D} \\
& +\frac{1}{6}\frac{\sum_{\zeta\eta\xi\omega}\mu^4_{\zeta\eta\xi\omega}\hat{b}_{D,\eta\xi\omega}\hat{f}_{z,\zeta}}{\hat{b}_D} - \frac{1}{48}\sum_{ijkqrs}\widehat{h}_{ijkq}\mu^6_{ijkqrs}\hat{f}_{z,rs} \\
& +\frac{1}{144}\sum_{ijkqrstw}\widehat{h}_{ijk}\widehat{h}_{qrs}\mu^8_{ijkqrstw}\hat{f}_{z,tw} - \frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\widehat{h}_{ijk}\mu^6_{ijk\zeta\eta\xi}\hat{f}_{z,\zeta\eta}\widehat{b}_{D,\xi}}{\widehat{b}_D} \\
& +\frac{1}{4}\frac{\sum_{\zeta\eta\xi\omega}\mu^4_{\zeta\eta\xi\omega}\hat{f}_{z,\zeta\eta}\widehat{b}_{D,\xi\omega}}{\widehat{b}_D}.
\end{aligned}$$

Let $B_{2,z}^2 = B_{21,z}^2 + B_{22,z}^2$, where

$$\begin{aligned}
B_{21,z}^2 \quad = \quad & -\frac{1}{120}\sum_{ijkqrs}\hat{h}_{ijkqr}\mu^6_{ijkqrs}\hat{f}_{z,s} + \frac{1}{144}\sum_{ijkqrstw}\hat{h}_{ijk}\hat{h}_{qrst}\mu^8_{ijkqrstw}\hat{f}_{z,w} \\
& -\frac{1}{1296}\sum_{ijkqrstwv\beta}\hat{h}_{ijk}\hat{h}_{qrs}\hat{h}_{twv}\mu^{10}_{ijkqrstwv\beta}\hat{f}_{z,\beta} - \frac{1}{24}\frac{\sum_{ijkqrs}\hat{h}_{ijkq}\mu^6_{ijkqrs}\hat{b}_{D,s}\hat{f}_{z,r}}{\hat{b}_D} \\
& +\frac{1}{72}\frac{\sum_{ijkqrstw}\hat{h}_{ijk}\hat{h}_{qrs}\mu^8_{ijkqrstw}\hat{b}_{D,w}\hat{f}_{z,t}}{\hat{b}_D} - \frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\hat{h}_{ijk}\mu^6_{ijk\zeta\eta\xi}\hat{b}_{D,\eta\xi}\hat{f}_{z,\zeta}}{\hat{b}_D} \\
& +\frac{1}{6}\frac{\sum_{\zeta\eta\xi\omega}\mu^4_{\zeta\eta\xi\omega}\hat{b}_{D,\eta\xi\omega}\hat{f}_{z,\zeta}}{\hat{b}_D},
\end{aligned}$$

$$\begin{aligned}
B_{22,z}^2 \quad = \quad & -\frac{1}{48}\sum_{ijkqrs}\widehat{h}_{ijkq}\mu^6_{ijkqrs}\hat{f}_{z,rs} \\
& +\frac{1}{144}\sum_{ijkqrstw}\widehat{h}_{ijk}\widehat{h}_{qrs}\mu^8_{ijkqrstw}\hat{f}_{z,tw} - \frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\widehat{h}_{ijk}\mu^6_{ijk\zeta\eta\xi}\hat{f}_{z,\zeta\eta}\widehat{b}_{D,\xi}}{\widehat{b}_D} \\
& +\frac{1}{4}\frac{\sum_{\zeta\eta\xi\omega}\mu^4_{\zeta\eta\xi\omega}\hat{f}_{z,\zeta\eta}\widehat{b}_{D,\xi\omega}}{\widehat{b}_D}.
\end{aligned}$$

Then, we rewrite them in the matrix form so that we have

$$B_2^2 = B_{21}^2 + B_{22}^2, \tag{71}$$

where

$$B_{21}^2 = \hat{\mathbf{f}}^{(1)} B_2^1 = \left(\hat{\theta}\otimes\mathbf{I}_P + \mathbf{I}_P\otimes\hat{\theta}\right) B_2^1 = vec\left(B_2^1\hat{\theta}' + \hat{\theta}B_2^{1\prime}\right). \tag{72}$$

By Li et al. (2017)

$$
\begin{aligned}
& B_{22}^{2} \qquad (73)\\
= \; & -\frac{1}{8}vec\left(\hat{h}^{(-2)}\right)\mathbf{tr}\left[\left(\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)'\right)\hat{h}^{(4)}\right]-\frac{1}{4}vec\left[\left(\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)'\right)\hat{h}^{(4)}\hat{h}^{(-2)}\right]\\
& -\frac{1}{4}vec\left[\hat{h}^{(-2)}\hat{h}^{(4)\prime}\left(\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)\right)\right]\\
& +\frac{1}{8}vec\left(\hat{h}^{(-2)}\right)\left[vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\right]\\
& +\frac{1}{12}vec\left(\hat{h}^{(-2)}\right)\mathbf{tr}\left[\hat{h}^{(3)\prime}\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\hat{h}^{(-2)}\right]\\
& +\frac{1}{4}vec\left[\left[\left(vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right)\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\hat{h}^{(-2)}\right]\\
& +\frac{1}{4}vec\left[\hat{h}^{(-2)}\hat{h}^{(3)\prime}\left[\left(vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right)'\otimes\hat{h}^{(-2)}\right]\right]\\
& +\frac{1}{4}vec\left[\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right]\\
& +\frac{1}{2}vec\left[\hat{h}^{(-2)}\hat{h}^{(3)\prime}\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\hat{h}^{(-2)}\right]-\frac{1}{2}vec\left[\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\frac{\hat{b}_{D}^{(1)}{}'}{\hat{b}_{D}}\hat{h}^{(-2)}\right]\\
& -\frac{1}{2}vec\left[\hat{h}^{(-2)}\frac{\hat{b}_{D}^{(1)}}{\hat{b}_{D}}vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right]-\frac{1}{2}vec\left[\left[\left(\hat{h}^{(-2)}\frac{\hat{b}_{D}^{(1)}}{\hat{b}_{D}}\right)'\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\hat{h}^{(-2)}\right]\\
& -\frac{1}{2}vec\left[\hat{h}^{(-2)}\hat{h}^{(3)\prime}\left[\left(\hat{h}^{(-2)}\frac{\hat{b}_{D}^{(1)}}{\hat{b}_{D}}\right)\otimes\hat{h}^{(-2)}\right]\right]-\frac{1}{2}vec\left(\hat{h}^{(-2)}\right)vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\frac{\hat{b}_{D}^{(1)}}{\hat{b}_{D}}\\
& +\frac{1}{2}vec\left(\hat{h}^{(-2)}\right)\mathbf{tr}\left[\hat{h}^{(-2)}\frac{\hat{b}_{D}^{(2)}}{\hat{b}_{D}}\right]+vec\left(\hat{h}^{(-2)}\frac{\hat{b}_{D}^{(2)}}{\hat{b}_{D}}\hat{h}^{(-2)}\right) \qquad (74)
\end{aligned}
$$

We can also get

$$B_{3}^{2}=B_{1}^{2}\times B_{4}^{1}=\left(vec\left(\hat{h}^{(-2)}\right)+\hat{\mathbf{f}}^{(1)}B_{1}^{1}\right)B_{4}^{1}, \qquad (75)$$

where

$$\hat{\mathbf{f}}^{(1)}B_{1}^{1}=vec\left(B_{1}^{1}\hat{\theta}'+\hat{\theta}B_{1}^{1\prime}\right).$$

Note that

$$
\begin{aligned}
\overline{\theta} &= \hat{\theta}+\frac{1}{n}B_{1}^{1}+\frac{1}{n^{2}}\left(B_{2}^{1}-B_{3}^{1}\right)+O_{p}\left(\frac{1}{n^{3}}\right)\\
&= \hat{\theta}+\frac{1}{n}B_{1}^{1}+\frac{1}{n^{2}}\left(B_{2}^{1}-B_{4}^{1}B_{1}^{1}\right)+O_{p}\left(\frac{1}{n^{3}}\right).
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
vec\left(\overline{\theta}\overline{\theta}'\right) &= vec\left(\hat{\theta}\hat{\theta}'\right)+\frac{1}{n}vec\left(\hat{\theta}B_{1}^{1\prime}+B_{1}^{1}\hat{\theta}'\right)\\
&\quad +\frac{1}{n^{2}}vec\left[\hat{\theta}\left(B_{2}^{1}-B_{4}^{1}B_{1}^{1}\right)'+\left(B_{2}^{1}-B_{4}^{1}B_{1}^{1}\right)\hat{\theta}'+B_{1}^{1}B_{1}^{1\prime}\right]+O_{p}\left(\frac{1}{n^{3}}\right).
\end{aligned}
$$

From (70), (71) and (75), we can show that

$$\begin{aligned}
&\frac{\int vec\left(\theta\theta'\right) b_D\left(\theta\right)\exp\left(-nh_n\left(\theta\right)\right)d\theta}{\int b_D\left(\theta\right)\exp\left(-nh_n\left(\theta\right)\right)d\theta}\\
= \;& vec\left(\hat{\theta}\hat{\theta}'\right)+\frac{1}{n}B_1^2+\frac{1}{n^2}\left(B_2^2-B_3^2\right)+O_p\left(\frac{1}{n^3}\right)\\
= \;& vec\left(\hat{\theta}\hat{\theta}'\right)+\frac{1}{n}\left[vec\left(\hat{h}^{(-2)}\right)+\hat{\mathbf{f}}^{(1)}B_1^1\right]\\
&+\frac{1}{n^2}\left[\hat{\mathbf{f}}^{(1)}B_2^1+B_{22}^2-B_4^1\left(vec\left(\hat{h}^{(-2)}\right)+\hat{\mathbf{f}}^{(1)}B_1^1\right)\right]+O_p\left(\frac{1}{n^3}\right)\\
= \;& vec\left(\hat{\theta}\hat{\theta}'\right)+\frac{1}{n}\left[vec\left(\hat{h}^{(-2)}\right)+vec\left(B_1^1\hat{\theta}'+\hat{\theta}B_1^{1\prime}\right)\right]\\
&+\frac{1}{n^2}\left[vec\left(B_2^1\hat{\theta}'+\hat{\theta}B_2^{1\prime}\right)+B_{22}^2-B_4^1\left(vec\left(\hat{h}^{(-2)}\right)+vec\left(B_1^1\hat{\theta}'+\hat{\theta}B_1^{1\prime}\right)\right)\right]+O_p\left(\frac{1}{n^3}\right)\\
= \;& vec\left(\hat{\theta}\hat{\theta}'\right)+\frac{1}{n}\left[vec\left(\hat{h}^{(-2)}\right)+vec\left(B_1^1\hat{\theta}'+\hat{\theta}B_1^{1\prime}\right)\right]\\
&+\frac{1}{n^2}\left[B_{22}^2+\hat{\theta}\left(B_2^1-B_4^1B_1^1\right)'+\left(B_2^1-B_4^1B_1^1\right)\hat{\theta}'-B_4^1vec\left(\hat{h}^{(-2)}\right)\right]+O_p\left(\frac{1}{n^3}\right).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\frac{\int vec\left[\left(\theta-\overline{\theta}\right)\left(\theta-\overline{\theta}\right)'\right] b_D\left(\theta\right)\exp\left(-nh_n\left(\theta\right)\right)d\theta}{\int b_D\left(\theta\right)\exp\left(-nh_n\left(\theta\right)\right)d\theta}\\
= \;& \frac{\int vec\left(\theta\theta'\right) b_D\left(\theta\right)\exp\left(-nh_n\left(\theta\right)\right)d\theta}{\int b_D\left(\theta\right)\exp\left(-nh_n\left(\theta\right)\right)d\theta}-vec\left(\overline{\theta}\overline{\theta}'\right)\\
= \;& \frac{1}{n}vec\left(\hat{h}^{(-2)}\right)+\frac{1}{n^2}\left[B_{22}^2-B_4^1vec\left(\hat{h}^{(-2)}\right)-vec\left(B_1^1B_1^{1\prime}\right)\right]+O_p\left(\frac{1}{n^3}\right).
\end{aligned}$$

Note that

$$\begin{aligned}
&vec\left(\hat{h}^{(-2)}\right)B_4^1\\
= \;& \frac{1}{2}vec\left(\hat{h}^{(-2)}\right)\mathbf{tr}\left[\hat{h}^{(-2)}\frac{\hat{b}_D^{(2)}}{\hat{b}_D}\right]-\frac{1}{2}vec\left(\hat{h}^{(-2)}\right)vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(-3)}\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D}\\
&+\frac{1}{8}vec\left(\hat{h}^{(-2)}\right)vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\\
&+\frac{1}{12}vec\left(\hat{h}^{(-2)}\right)vec\left(\hat{h}^{(3)}\right)'\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]vec\left(\hat{h}^{(3)}\right)\\
&-\frac{1}{8}vec\left(\hat{h}^{(-2)}\right)\mathbf{tr}\left[\left[\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)\right]\hat{h}^{(4)\prime}\right],
\end{aligned}$$

and that

$$\begin{aligned}
B_1^1B_1^{1\prime} &= \left[\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D}-\frac{1}{2}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\right]\left[\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D}-\frac{1}{2}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\right]'\\
&= \hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D}\frac{\hat{b}_D^{(1)\prime}}{\hat{b}_D}\hat{h}^{(-2)}-\frac{1}{2}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\frac{\hat{b}_D^{(1)\prime}}{\hat{b}_D}\hat{h}^{(-2)}
\end{aligned}$$

$$-\frac{1}{2}\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D}vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}+\frac{1}{4}\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}.$$

We have

$$\begin{aligned}
& B_{22}^2-B_4^1vec\left(\hat{h}^{(-2)}\right)\\
= & -\frac{1}{4}vec\left[\left(\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)'\right)\hat{h}^{(4)}\hat{h}^{(-2)}\right]-\frac{1}{4}vec\left[\hat{h}^{(-2)}\hat{h}^{(4)\prime}\left(\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)\right)\right]\\
& +\frac{1}{4}vec\left[\left[\left(vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right)\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\hat{h}^{(-2)}\right]\\
& +\frac{1}{4}vec\left[\hat{h}^{(-2)}\hat{h}^{(3)\prime}\left[\left(vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right)'\otimes\hat{h}^{(-2)}\right]\right]\\
& +\frac{1}{4}vec\left[\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right]\\
& +\frac{1}{2}vec\left[\hat{h}^{(-2)}\hat{h}^{(3)\prime}\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\hat{h}^{(-2)}\right]-\frac{1}{2}vec\left[\hat{h}^{(-2)}\hat{h}^{(3)\prime}vec\left(\hat{h}^{(-2)}\right)\frac{\hat{b}_D^{(1)\prime}}{\hat{b}_D}\hat{h}^{(-2)}\right]\\
& -\frac{1}{2}vec\left[\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D}vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right]-\frac{1}{2}vec\left[\left[\left(\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D}\right)'\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\hat{h}^{(-2)}\right]\\
& -\frac{1}{2}vec\left[\hat{h}^{(-2)}\hat{h}^{(3)\prime}\left[\left(\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D}\right)\otimes\hat{h}^{(-2)}\right]\right]+vec\left(\hat{h}^{(-2)}\frac{\hat{b}_D^{(2)}}{\hat{b}_D}\hat{h}^{(-2)}\right).
\end{aligned}$$

We can further decompose $B_{22}^2-B_4^1vec\left(\hat{h}^{(-2)}\right)-vec\left(B_1^1B_1^{1\prime}\right)$ as

$$B_{22}^2-B_4^1vec\left(\hat{h}^{(-2)}\right)-vec\left(B_1^1B_1^{1\prime}\right)=F_1+F_2,$$

where

$$\begin{aligned}
F_1 = & -\frac{1}{4}vec\left[\left(\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)'\right)\hat{h}^{(4)}\hat{h}^{(-2)}\right]-\frac{1}{4}vec\left[\hat{h}^{(-2)}\hat{h}^{(4)\prime}\left(\hat{h}^{(-2)}\otimes vec\left(\hat{h}^{(-2)}\right)\right)\right]\\
& +\frac{1}{4}vec\left[\hat{h}^{(-2)}\hat{h}^{(3)\prime}\left[\left(vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right)'\otimes\hat{h}^{(-2)}\right]\right]\\
& +\frac{1}{4}vec\left[\left[\left(vec\left(\hat{h}^{(-2)}\right)'\hat{h}^{(3)}\hat{h}^{(-2)}\right)\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\hat{h}^{(-2)}\right]\\
& +\frac{1}{2}vec\left[\hat{h}^{(-2)}\hat{h}^{(3)\prime}\left[\hat{h}^{(-2)}\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\hat{h}^{(-2)}\right], \qquad (76)
\end{aligned}$$

$$\begin{aligned}
F_2 = & -\frac{1}{2}vec\left[\left[\left(\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D}\right)'\otimes\hat{h}^{(-2)}\right]\hat{h}^{(3)}\hat{h}^{(-2)}\right]-\frac{1}{2}vec\left[\hat{h}^{(-2)}\hat{h}^{(3)\prime}\left[\left(\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D}\right)\otimes\hat{h}^{(-2)}\right]\right]\\
& +vec\left(\hat{h}^{(-2)}\frac{\hat{b}_D^{(2)}}{\hat{b}_D}\hat{h}^{(-2)}\right)-vec\left(\hat{h}^{(-2)}\frac{\hat{b}_D^{(1)}}{\hat{b}_D}\frac{\hat{b}_D^{(1)\prime}}{\hat{b}_D}\hat{h}^{(-2)}\right). \qquad (77)
\end{aligned}$$

■

6.3 Proof of Theorem 3.1

It is noted that $h_n(\theta) = -\bar{l}_n(\theta) = -\frac{1}{n}\sum_{t=1}^{n} l_t(\theta)$, $b_D(\theta) = p(\theta)$, $\pi(\theta) = \ln p(\theta)$ and $\bar{\mathbf{H}}_n^{(j)}(\theta) = \frac{1}{n}\sum_{t=1}^{n} l_t^{(j)}(\theta) = \bar{l}_n^{(j)}(\theta)$ for $j = 3, 4,$. Thus, according to Lemma 3.2, we have

$$\begin{aligned}\overline{\theta} &= \frac{\int \theta p(\theta)\exp(-nh_n(\theta))\,d\theta}{\int p(\theta)\exp(-nh_n(\theta))\,d\theta} = \hat{\theta} - \frac{1}{n}\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}} \\ &\quad + \frac{1}{2n}\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)' vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right) + O_p\left(\frac{1}{n^2}\right), \end{aligned} \tag{78}$$

and

$$\begin{aligned} vec\left(V\left(\overline{\theta}\right)\right) &= \frac{\int vec\left[\left(\theta-\overline{\theta}\right)\left(\theta-\overline{\theta}\right)'\right] p(\theta)\exp(-nh_n(\theta))\,d\theta}{\int p(\theta)\exp(-nh_n(\theta))\,d\theta} \\ &= -\frac{1}{n}vec\left(\hat{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right) + \frac{1}{n^2}F_1 + \frac{1}{n^2}F_2 + O_p\left(\frac{1}{n^3}\right), \end{aligned} \tag{79}$$

where

$$\begin{aligned} F_1 &= -\frac{1}{4}vec\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\otimes vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right] \\ &\quad -\frac{1}{4}vec\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(4)}\left(\hat{\theta}\right)'\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\otimes vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)\right)\right] \\ &\quad +\frac{1}{4}vec\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\left(vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\right] \\ &\quad +\frac{1}{4}vec\left[\left[\left(vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right] \\ &\quad +\frac{1}{2}vec\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right] \end{aligned} \tag{80}$$

$$\begin{aligned} F_2 &= -\frac{1}{2}vec\left[\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right)'\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right] \\ &\quad -\frac{1}{2}vec\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\right] \\ &\quad +vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(2)}}{\hat{p}}\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right) - vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\frac{\hat{p}^{(1)}}{\hat{p}}'\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right) \end{aligned} \tag{81}$$

From (78), by the Taylor expansion of $vec\left(\bar{\mathbf{H}}_n\left(\overline{\theta}\right)\right)$ at $\hat{\theta}$, we have

$$vec\left(\bar{\mathbf{H}}_n\left(\overline{\theta}\right)\right) = vec\left[\bar{\mathbf{H}}_n\left(\hat{\theta}\right) + \bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\left(\overline{\theta}-\hat{\theta}\right)\right] + O_p\left(\frac{1}{n^2}\right). \tag{82}$$

Hence, we get

$$\begin{aligned} P_L &= \mathbf{tr}\left[-n\bar{\mathbf{H}}_n\left(\overline{\theta}\right)V\left(\overline{\theta}\right)\right] = -nvec\left(\bar{\mathbf{H}}_n\left(\overline{\theta}\right)\right)'vec\left(V\left(\overline{\theta}\right)\right) \\ &= -nvec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)\right)'vec\left(V\left(\overline{\theta}\right)\right) - nvec\left(\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\left(\overline{\theta}-\hat{\theta}\right)\right)'vec\left(V\left(\overline{\theta}\right)\right) \end{aligned}$$

$$-nvec\left(V\left(\overline{\theta}\right)\right)O_p\left(\frac{1}{n^2}\right) \tag{83}$$

By (67), (68), and (70), we can have

$$\begin{aligned} & nvec\left(\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\left(\overline{\theta}-\hat{\theta}\right)\right)' vec\left(V\left(\overline{\theta}\right)\right) \qquad (84)\\ = \;& vec\left[\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\left(\overline{\theta}-\hat{\theta}\right)\right]' vec\left(-\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)+O_p(\frac{1}{n^2})\\ = \;& \frac{1}{n}vec\left[\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\left(-\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}+\frac{1}{2}\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)' vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)\right)\right]'\\ & \left[vec\left(-\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)\right]+O_p(\frac{1}{n^2})\\ = \;& \frac{1}{n}\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)' vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)\\ & -\frac{1}{2n}vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)' vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)+O_p(\frac{1}{n^2}) \end{aligned}$$

where

$$\begin{aligned} & vec\left[\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right]' vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right) \qquad (85)\\ = \;& vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\\ = \;& vec\left(\mathbf{I}_P\times\mathbf{I}_P\times\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\\ = \;& vec\left(\mathbf{I}_P\right)' vec\left[\left[\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\otimes\mathbf{I}_P\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right]\\ = \;& vec\left(\mathbf{I}_P\right)'\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right)'\otimes\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\otimes\mathbf{I}_P\right]vec\left(\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\right)\\ = \;& \mathbf{tr}\left[\mathbf{I}_P\times\mathbf{I}_P\times\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\times\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right)\otimes\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right]\right]\\ = \;& \mathbf{tr}\left[\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right)\otimes\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right]\right], \end{aligned}$$

by (50) and (51). For the same reason

$$\begin{aligned} & vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)' vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right) \qquad (86)\\ = \;& \mathbf{tr}\left[\left[\left(vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\right]. \end{aligned}$$

Then from (84), (85) and (86), we have

$$nvec\left(\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\left(\overline{\theta}-\hat{\theta}\right)\right)' vec\left(V\left(\overline{\theta}\right)\right) \tag{87}$$

$$
\begin{aligned}
&= \frac{1}{n}\mathbf{tr}\left[\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right)\otimes\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right]\right] \\
&\quad -\frac{1}{2n}\mathbf{tr}\left[\left[\left(vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\right]+O_p(\frac{1}{n^2}).
\end{aligned}
$$

And note that

$$
\begin{aligned}
& nvec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)\right)'vec\left(V\left(\overline{\theta}\right)\right) \\
= \;& vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)\right)'\left[-vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)+\frac{1}{n}F_1+\frac{1}{n}F_2\right]+O_p\left(\frac{1}{n^2}\right) \\
= \;& -P+\frac{1}{n}vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)\right)'F_1+\frac{1}{n}vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)\right)'F_2+O_p\left(\frac{1}{n^2}\right). \qquad (88)
\end{aligned}
$$

Furthermore, from (80) and (81), we have

$$
\begin{aligned}
& vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)\right)'F_1 \\
= \;& -\frac{1}{2}\mathbf{tr}\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\otimes vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\hat{\theta}\right)\right] \\
& +\frac{1}{2}\mathbf{tr}\left[\left[\left(vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\right] \\
& +\frac{1}{2}\mathbf{tr}\left[\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right] \qquad (89)
\end{aligned}
$$

$$
\begin{aligned}
& vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)\right)'F_2 \\
= \;& -\mathbf{tr}\left[\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\right] \\
& +\mathbf{tr}\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(2)}}{\hat{p}}\right]-\mathbf{tr}\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\frac{\hat{p}^{(1)\prime}}{\hat{p}}\right] \qquad (90)
\end{aligned}
$$

Hence, from (89) and (90), we have

$$
\begin{aligned}
& nvec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)\right)'vec\left(V\left(\overline{\theta}\right)\right) \qquad (91) \\
= \;& -P-\frac{1}{2n}\mathbf{tr}\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\otimes vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\hat{\theta}\right)\right] \\
& +\frac{1}{2n}\mathbf{tr}\left[\left[\left(vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\right] \\
& +\frac{1}{2n}\mathbf{tr}\left[\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right] \\
& -\mathbf{tr}\left[\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\right] \\
& +\mathbf{tr}\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(2)}}{\hat{p}}\right]-\mathbf{tr}\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\frac{\hat{p}^{(1)\prime}}{\hat{p}}\right].
\end{aligned}
$$

Then, from (83) and (91), we have

$$P_L = P + \frac{1}{n}C_1 + \frac{1}{n}C_2 + O_p\left(\frac{1}{n^2}\right),$$

where

$$\begin{aligned} C_1 &= \frac{1}{2}\mathbf{tr}\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right) \otimes vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\hat{\theta}\right)\right] \\ &\quad -\frac{1}{2}\mathbf{tr}\left[\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right) \otimes \bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right] \end{aligned}$$

$$C_2 = -\mathbf{tr}\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(2)}}{\hat{p}}\right] + \mathbf{tr}\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\frac{\hat{p}^{(1)\prime}}{\hat{p}}\right] = -\mathbf{tr}\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\pi^{(2)}\left(\hat{\theta}\right)\right].$$

We can rewrite C_1 and C_2 as

$$C_1 = \frac{1}{2}C_{11} - \frac{1}{2}C_{12}, C_2 = -C_{22},$$

where

$$C_{11} = \mathbf{tr}\left[\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1} \otimes vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\hat{\theta}\right)\right],$$

$$\begin{aligned} C_{12} &= \mathbf{tr}\left[\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right) \otimes \bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right] \\ &= vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)' vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right), \end{aligned}$$

$$C_{22} = \mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\pi^{(2)}\left(\hat{\theta}\right)\right], C_{23} = \pi^{(1)}\left(\hat{\theta}\right)'\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\pi^{(1)}\left(\hat{\theta}\right).$$

And from Li et al. (2017)

$$\ln p\left(\mathbf{y}|\overline{\theta}\right) = \ln p\left(\mathbf{y}|\hat{\theta}\right) - \frac{1}{2n}C_{21} + \frac{1}{2n}C_{23} + \frac{1}{8n}C_{12} + O_p\left(n^{-2}\right) \tag{92}$$

where

$$\begin{aligned} C_{21} &= \mathbf{tr}\left[\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\pi^{(1)}\left(\hat{\theta}\right)\right) \otimes \bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\right] \\ &= \pi^{(1)}\left(\hat{\theta}\right)'\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)' vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right). \end{aligned}$$

Hence

$$\begin{aligned} \mathrm{DIC}_L &= -2\ln p\left(\mathbf{y}|\overline{\theta}\right) + 2P_L \\ &= -2\ln p\left(\mathbf{y}|\hat{\theta}\right) + \frac{1}{n}C_{21} - \frac{1}{n}C_{23} - \frac{1}{4n}C_{12} + 2P + \frac{2}{n}C_1 + \frac{2}{n}C_2 + O_p\left(\frac{1}{n^2}\right) \\ &= \mathrm{AIC} + \frac{1}{n}D_1 + \frac{1}{n}D_2 + O_p\left(\frac{1}{n^2}\right), \end{aligned}$$

where

$$D_1 = C_{11} + \frac{5}{4}C_{12},$$
$$D_2 = C_{21} - 2C_{22} - C_{23}.$$

6.4 Proof of Theorem 4.1

By the second-order Taylor expansion of $\mathbf{s}_t\left(\bar{\theta}\right)$ at $\hat{\theta}$, $\bar{\mathbf{\Omega}}_n\left(\bar{\theta}\right)$ can be written as

$$\begin{aligned}
\bar{\mathbf{\Omega}}_n\left(\bar{\theta}\right) &= \frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\left[\mathbf{s}_t\left(\hat{\theta}\right)+l_t^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)+\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)'\right)l_t^{(3)}\left(\tilde{\theta}_t\right)\left(\bar{\theta}-\hat{\theta}\right)\right] \\
&\quad\times\left[\mathbf{s}_\tau\left(\hat{\theta}\right)+l_\tau^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)+\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)'\right)l_\tau^{(3)}\left(\tilde{\theta}_\tau\right)\left(\bar{\theta}-\hat{\theta}\right)\right]' k\left(\frac{t-\tau}{\gamma_n}\right) \\
&= \frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\mathbf{s}_t\left(\hat{\theta}\right)\mathbf{s}_\tau\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right)+\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\mathbf{s}_t\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(2)}\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right) \\
&\quad+\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}l_t^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)\mathbf{s}_\tau\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right) \\
&\quad+\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}l_t^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(2)}\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right) \\
&\quad+\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\mathbf{s}_t\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(3)}\left(\tilde{\theta}_\tau\right)'\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)\right)k\left(\frac{t-\tau}{\gamma_n}\right) \\
&\quad+\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}l_t^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(3)}\left(\tilde{\theta}_\tau\right)'\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)\right)k\left(\frac{t-\tau}{\gamma_n}\right) \\
&\quad+\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)'\right)l_t^{(3)}\left(\tilde{\theta}_t\right)\left(\bar{\theta}-\hat{\theta}\right)\mathbf{s}_\tau\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right) \\
&\quad+\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)'\right)l_t^{(3)}\left(\tilde{\theta}_t\right)\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(2)}\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right) \\
&\quad+\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)'\right)l_t^{(3)}\left(\tilde{\theta}_t\right)\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(3)}\left(\vec{\theta}_\tau\right)'\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)\right)k\left(\frac{t-\tau}{\gamma_n}\right),
\end{aligned}$$

where both $\tilde{\theta}_t$ and $\tilde{\theta}_\tau$ lie between $\bar{\theta}$ and $\hat{\theta}$ for all t and τ. We consider the stochastic order of each term. For simplicity, we first consider the terms with order greater than or equal to $O_p\left(\gamma_n/n^2\right)$ which are the fourth to ninth terms. Without of loss of generality, we will analyze the fifth and sixth terms only.

For the fifth term, we can rewrite it as

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\mathbf{s}_t\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(3)}\left(\tilde{\theta}_\tau\right)'\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)\right)k\left(\frac{t-\tau}{\gamma_n}\right) \qquad (93)\\
= &\left[\frac{1}{n^2}\sum_{t=1}^{n}\frac{1}{n}\sum_{\tau=1}^{n}\mathbf{s}_t\left(\hat{\theta}\right)n\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(3)}\left(\tilde{\theta}_\tau\right)' k\left(\frac{t-\tau}{\gamma_n}\right)\right]\left(I_p\otimes n\left(\bar{\theta}-\hat{\theta}\right)\right) \\
= &\left[\begin{array}{c}\frac{1}{n^2}\sum_{j=0}^{n-1}k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=j+1}^{n}\mathbf{s}_t\left(\hat{\theta}\right)n\left(\bar{\theta}-\hat{\theta}\right)' l_{t-j}^{(3)}\left(\tilde{\theta}_{t-j}\right)' \\ +\frac{1}{n^2}\sum_{j=-n+1}^{-1}k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=-j+1}^{n}\mathbf{s}_{t+j}\left(\hat{\theta}\right)n\left(\bar{\theta}-\hat{\theta}\right)' l_t^{(3)}\left(\tilde{\theta}_t\right)'\end{array}\right]\times\left(I_p\otimes n\left(\bar{\theta}-\hat{\theta}\right)\right).
\end{aligned}$$

We have $I_p\otimes n\left(\bar{\theta}-\hat{\theta}\right)=O_p\left(1\right)$ by Lemma 3.2 , then we need to consider the order of the

first term in (93), that is

$$
\begin{aligned}
& vec\left[\frac{1}{n^2}\sum_{t=1}^{n}\frac{1}{n}\sum_{\tau=1}^{n}\mathbf{s}_t\left(\hat{\theta}\right)n\left(\bar{\theta}-\hat{\theta}\right)'l_{\tau}^{(3)}\left(\tilde{\theta}_{\tau}\right)'k\left(\frac{t-\tau}{\gamma_n}\right)\right] \qquad (94)\\
= & \left[\begin{array}{c}\frac{1}{n^2}\sum_{j=0}^{n-1}k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=j+1}^{n}\left[l_{t-j}^{(3)}\left(\tilde{\theta}_{t-j}\right)\otimes\mathbf{s}_t\left(\hat{\theta}\right)\right]\\ +\frac{1}{n^2}\sum_{j=-n+1}^{-1}k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=-j+1}^{n}\left[l_{t}^{(3)}\left(\tilde{\theta}_{t}\right)\otimes\mathbf{s}_{t+j}\left(\hat{\theta}\right)\right]\end{array}\right]\times vec\left(n\left(\bar{\theta}-\hat{\theta}\right)\right).
\end{aligned}
$$

By the Minkowski inequality

$$
\begin{aligned}
& \left\|\begin{array}{c}\frac{1}{n^2}\sum_{j=0}^{n-1}k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=j+1}^{n}\left[l_{t-j}^{(3)}\left(\tilde{\theta}_{t-j}\right)\otimes\mathbf{s}_t\left(\hat{\theta}\right)\right]\\ +\frac{1}{n^2}\sum_{j=-n+1}^{-1}k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=-j+1}^{n}\left[l_{t}^{(3)}\left(\tilde{\theta}_{t}\right)\otimes\mathbf{s}_{t+j}\left(\hat{\theta}\right)\right]\end{array}\right\| \qquad (95)\\
\leq & \frac{1}{n^2}\sum_{j=0}^{n-1}\left|k\left(\frac{j}{\gamma_n}\right)\right|\frac{1}{n}\left\|\sum_{t=j+1}^{n}l_{t-j}^{(3)}\left(\tilde{\theta}_{t-j}\right)\otimes\mathbf{s}_t\left(\hat{\theta}\right)\right\|+\\
& \frac{1}{n^2}\sum_{j=-n+1}^{-1}\left|k\left(\frac{j}{\gamma_n}\right)\right|\frac{1}{n}\left\|\sum_{t=-j+1}^{n}l_{t}^{(3)}\left(\tilde{\theta}_{t}\right)\otimes\mathbf{s}_{t+j}\left(\hat{\theta}\right)\right\|.
\end{aligned}
$$

Following Gallant and White (1988), we consider each element of $l_{t-j}^{(3)}\left(\tilde{\theta}_{t-j}\right)\otimes\mathbf{s}_t\left(\hat{\theta}\right)$ which can be expressed as the product of (uv)th element of $l_{t-j}^{(3)}\left(\tilde{\theta}_{t-j}\right)$ and wth element of $s_t\left(\hat{\theta}\right)$ for. It is denoted by $l_{t-j,uv}^{(3)}\left(\tilde{\theta}_{t-j}\right)s_{t,w}\left(\hat{\theta}\right)$ for each $j\geq 0$. Then we have

$$
\begin{aligned}
\left\|\frac{1}{n}\sum_{t=j+1}^{n}l_{t-j,uv}^{(3)}\left(\tilde{\theta}_t\right)\mathbf{s}_{t,w}\left(\hat{\theta}\right)\right\| & \leq \left(\frac{1}{n}\sum_{t=j+1}^{n}\left\|l_{t-j,uv}^{(3)}\left(\tilde{\theta}_t\right)\right\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{t=j+1}^{n}\left\|\mathbf{s}_{t,w}\left(\hat{\theta}\right)\right\|^2\right)^{1/2}\\
& \leq \left(\frac{1}{n}\sum_{t=j+1}^{n}M_t^2\right)^{1/2}\left(\frac{1}{n}\sum_{t=j+1}^{n}M_t^2\right)^{1/2}\leq\frac{1}{n}\sum_{t=1}^{n}M_t^2,
\end{aligned}
$$

for each j and each element of $\frac{1}{n}\sum_{t=j+1}^{n}l_{t-j}^{(3)}\left(\tilde{\theta}_t\right)\mathbf{s}_t\left(\hat{\theta}\right)$. The first inequality follows the Cauchy-Schwarz inequality and the second inequality is due to Assumption 5. Then

$$
\left\|\frac{1}{n}\sum_{t=j+1}^{n}l_{t-j}^{(3)}\left(\tilde{\theta}_{t-j}\right)\otimes\mathbf{s}_t\left(\hat{\theta}\right)\right\|^2\leq P^3\left(\frac{1}{n}\sum_{t=1}^{n}M_t^2\right)^2,
$$

for each $j\geq 0$ since there are P^3 elements in $\frac{1}{n}\sum_{t=j+1}^{n}l_t^{(3)}\left(\tilde{\theta}_t\right)\mathbf{s}_{t-j}\left(\hat{\theta}\right)$. And also by Assumption 5, $\sup_t E\left(M_t^2\right)\leq M^2<\infty$. Similar to Andrews (1991), by Markov's inequality, $\frac{1}{n}\sum_{t=1}^{n}M_t^2=O_p\left(1\right)$, we have $\left\|\frac{1}{n}\sum_{t=j+1}^{n}l_t^{(3)}\left(\tilde{\theta}_t\right)\mathbf{s}_{t-j}\left(\hat{\theta}\right)\right\|=O_p\left(1\right)$ for each $j\geq 0$. Hence, we have

$$
\sup_{0\leq j\leq n-1}\left\|\frac{1}{n}\sum_{t=j+1}^{n}l_{t-j}^{(3)}\left(\tilde{\theta}_{t-j}\right)\otimes\mathbf{s}_t\left(\hat{\theta}\right)\right\|=O_p\left(1\right), \qquad (96)
$$

and

$$\sup_{-n+1\leq j\leq -1}\left\|\frac{1}{n}\sum_{t=-j+1}^{n}l_t^{(3)}\left(\tilde{\theta}_t\right)\otimes \mathbf{s}_{t+j}\left(\hat{\theta}\right)\right\| = O_p\left(1\right). \tag{97}$$

By (95), (96) and (97), we have

$$\begin{aligned}
&\frac{n^2}{\gamma_n}\left\| \begin{array}{c} \frac{1}{n^2}\sum_{j=0}^{n-1}k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=j+1}^{n}\left[l_{t-j}^{(3)}\left(\tilde{\theta}_{t-j}\right)\otimes \mathbf{s}_t\left(\hat{\theta}\right)\right] \\ +\frac{1}{n^2}\sum_{j=-n+1}^{-1}k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=-j+1}^{n}\left[l_t^{(3)}\left(\theta\right)\otimes \mathbf{s}_{t+j}\left(\theta\right)\right] \end{array}\right\| \\
\leq\ & \frac{1}{\gamma_n}\sum_{j=-n+1}^{n-1}\left|k\left(\frac{j}{\gamma_n}\right)\right| \\
&\times\max\left\{\sup_{0\leq j\leq n-1}\left\|\frac{1}{n}\sum_{t=j+1}^{n}l_t^{(3)}\left(\theta\right)\otimes \mathbf{s}_{t+j}\left(\theta\right)\right\|,\sup_{-n+1\leq j\leq -1}\left\|\frac{1}{n}\sum_{t=-j+1}^{n}l_t^{(3)}\left(\theta\right)\otimes \mathbf{s}_{t+j}\left(\theta\right)\right\|\right\} \\
=\ & O_p\left(1\right),
\end{aligned} \tag{98}$$

by the Minkowski inequality and Assumption 11. Then from (93), (94) and (98), we get

$$\frac{n^2}{\gamma_n}\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\mathbf{s}_t\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)'l_\tau^{(3)}\left(\tilde{\theta}_\tau\right)'\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)\right)k\left(\frac{t-\tau}{\gamma_n}\right)=O_p\left(1\right). \tag{99}$$

Similarly, for the sixth term, we have

$$\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}l_t^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)'l_\tau^{(3)}\left(\tilde{\theta}_\tau\right)'\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)\right)k\left(\frac{t-\tau}{\gamma_n}\right)=CC_1\times n\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)\right), \tag{100}$$

where

$$\begin{aligned}
CC_1 &= \frac{1}{n}\sum_{t=1}^{n}\frac{1}{n}\sum_{\tau=1}^{n}l_t^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)'l_\tau^{(3)}\left(\tilde{\theta}_\tau\right)'k\left(\frac{t-\tau}{\gamma_n}\right) \\
&= \frac{1}{n}\sum_{j=-0}^{n-1}k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=j+1}^{n}l_t^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)'l_{t-j}^{(3)}\left(\tilde{\theta}_{t-j}\right)' \\
&\quad +\frac{1}{n}\sum_{j=-n+1}^{-1}k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=-j+1}^{n}l_{t+j}^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)'l_t^{(3)}\left(\tilde{\theta}_t\right)',
\end{aligned}$$

and $n\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)\right)=O_p\left(1\right)$. Taking vectorization of CC_1, we have

$$vec(CC_1)=CC_2\times vec\left[n^2\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)'\right], \tag{101}$$

where

$$CC_2 = \frac{1}{n}\sum_{j=-0}^{n-1}k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=j+1}^{n}\left[l_{t-j}^{(3)}\left(\tilde{\theta}_{t-j}\right)\otimes l_t^{(2)}\left(\hat{\theta}\right)\right] \tag{102}$$

$$+\frac{1}{n}\sum_{j=-n+1}^{-1} k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=-j+1}^{n}\left[l_t^{(3)}\left(\tilde{\theta}_t\right)\otimes l_{t+j}^{(2)}\left(\hat{\theta}\right)\right],$$

and $vec\left[n^2\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)'\right]=O_p(1)$ by Lemma 3.2. Similar to the proof of (96) and (97), we have

$$\sup_{0\le j\le n-1}\left\|\frac{1}{n}\sum_{t=j+1}^{n} l_{t-j}^{(3)}\left(\tilde{\theta}_{t-j}\right)\otimes l_t^{(2)}\left(\hat{\theta}\right)\right\|=O_p(1), \tag{103}$$

and

$$\sup_{-n+1\le j\le -1}\left\|\frac{1}{n}\sum_{t=-j+1}^{n} l_t^{(3)}\left(\tilde{\theta}_t\right)\otimes l_{t+j}^{(2)}\left(\hat{\theta}\right)\right\|=O_p(1), \tag{104}$$

by Assumption 5. Hence, by (102), (103) and (104), we have

$$\begin{aligned}
&\frac{n^3}{\gamma_n}\|CC_2\|\\
\le\ &\frac{1}{\gamma_n}\sum_{j=-n+1}^{n-1}\left|k\left(\frac{j}{\gamma_n}\right)\right|\\
&\times\max\left\{\sup_{0\le j\le n-1}\left\|\frac{1}{n}\sum_{t=j+1}^{n} l_{t-j}^{(3)}\left(\tilde{\theta}_{t-j}\right)\otimes l_t^{(2)}\left(\hat{\theta}\right)\right\|,\ \sup_{-n+1\le j\le -1}\left\|\frac{1}{n}\sum_{t=-j+1}^{n} l_t^{(3)}\left(\tilde{\theta}_t\right)\otimes l_{t+j}^{(2)}\left(\hat{\theta}\right)\right\|\right\}\\
=\ &O_p(1).
\end{aligned} \tag{105}$$

Hence by (100), (101) and (105), we can get

$$\frac{n^3}{\gamma_n}\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n} l_t^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(3)}\left(\tilde{\theta}_\tau\right)'\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)\right)k\left(\frac{t-\tau}{\gamma_n}\right)=O_p(1). \tag{106}$$

In the same way, we can obtain the order for the fourth, seventh to ninth terms as

$$\frac{n^2}{\gamma_n}\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n} l_t^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(2)}\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right)=O_p(1), \tag{107}$$

$$\frac{n^2}{\gamma_n}\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)'\right) l_t^{(3)}\left(\tilde{\theta}_t\right)\left(\bar{\theta}-\hat{\theta}\right)\mathbf{s}_\tau\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right)=O_p(1), \tag{108}$$

$$\frac{n^3}{\gamma_n}\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)'\right) l_t^{(3)}\left(\tilde{\theta}_t\right)\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(2)}\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right)=O_p(1), \tag{109}$$

$$\frac{n^4}{\gamma_n}\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)'\right) l_t^{(3)}\left(\tilde{\theta}_t\right)\left(\bar{\theta}-\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(3)}\left(\tilde{\theta}_\tau\right)'\left(I_p\otimes\left(\bar{\theta}-\hat{\theta}\right)\right)k\left(\frac{t-\tau}{\gamma_n}\right)=O_p(1). \tag{110}$$

From (107), (99), (106), (108), (109) and (110), we have

$$\bar{\mathbf{\Omega}}_n(\bar{\theta}) = \bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)+\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\mathbf{s}_t\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(2)}\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right)$$

$$+\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n} l_t^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)\mathbf{s}_\tau\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right)+O_p\left(\frac{\gamma_n}{n^2}\right). \tag{111}$$

In (111), the second term can be written as

$$\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\mathbf{s}_t\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(2)}\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right)=\frac{\gamma_n}{n}W_1 \tag{112}$$

where

$$W_1=\frac{1}{\gamma_n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\mathbf{s}_t\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)' l_\tau^{(2)}\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right). \tag{113}$$

We have

$$\begin{aligned} W_1 &= \frac{1}{\gamma_n}\sum_{j=0}^{n-1} k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=j+1}^{n}\mathbf{s}_t\left(\hat{\theta}\right)n\left(\bar{\theta}-\hat{\theta}\right)' l_{t-j}^{(2)}\left(\hat{\theta}\right)' \\ &\quad +\frac{1}{\gamma_n}\sum_{j=-n+1}^{-1} k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=-j+1}^{n}\mathbf{s}_{t+j}\left(\hat{\theta}\right)n\left(\bar{\theta}-\hat{\theta}\right)' l_t^{(2)}\left(\hat{\theta}\right)'. \end{aligned}$$

Vectorization of W_1 is

$$vec\left(W_1\right)=W_{11}n\left(\bar{\theta}-\hat{\theta}\right),$$

where $n\left(\bar{\theta}-\hat{\theta}\right)=O_p\left(1\right)$ by Lemma 3.2 and

$$\begin{aligned} W_{11} &= \frac{1}{\gamma_n}\sum_{j=0}^{n-1} k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=j+1}^{n}\left[l_{t-j}^{(2)}\left(\hat{\theta}\right)\otimes\mathbf{s}_t\left(\hat{\theta}\right)\right] \\ &\quad +\frac{1}{\gamma_n}\sum_{j=-n+1}^{-1} k\left(\frac{j}{\gamma_n}\right)\frac{1}{n}\sum_{t=-j+1}^{n}\left[l_t^{(2)}\left(\hat{\theta}\right)\otimes\mathbf{s}_{t+j}\left(\hat{\theta}\right)\right]. \end{aligned}$$

Similar to (96) and (97), we can prove that

$$\sup_{0\le j\le n-1}\left\|\frac{1}{n}\sum_{t=j+1}^{n} l_{t-j}^{(2)}\left(\hat{\theta}\right)\otimes\mathbf{s}_t\left(\hat{\theta}\right)\right\|=O_p\left(1\right), \tag{114}$$

and

$$\sup_{-n+1\le j\le -1}\left\|\frac{1}{n}\sum_{t=-j+1}^{n} l_t^{(2)}\left(\hat{\theta}\right)\otimes\mathbf{s}_{t+j}\left(\hat{\theta}\right)\right\|=O_p\left(1\right), \tag{115}$$

by Assumption 5. Hence, by (114) and (115) and Assumption 11, we can get $W_{11}=O_p\left(1\right)$ and $W_1=O_p\left(1\right)$.

Similarly, the third term of (111) can be written as

$$\frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n} l_t^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)\mathbf{s}_\tau\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right)=\frac{\gamma_n}{n}W_2, \tag{116}$$

where

$$W_2 = \frac{1}{\gamma_n}\sum_{t=1}^{n}\sum_{\tau=1}^{n} l_t^{(2)}\left(\hat{\theta}\right)\left(\bar{\theta}-\hat{\theta}\right)\mathbf{s}_\tau\left(\hat{\theta}\right)' k\left(\frac{t-\tau}{\gamma_n}\right), \tag{117}$$

and $W_2 = O_p(1)$. By (112) and (116), we can rewrite (111) as

$$\bar{\mathbf{\Omega}}_n\left(\bar{\theta}\right) = \bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right) + \frac{\gamma_n}{n}\left(W_1 + W_2\right) + O_p\left(\frac{\gamma_n}{n^2}\right), \tag{118}$$

which proves (32) in Theorem 4.1.

For $vec\left(\bar{\mathbf{\Omega}}_n\left(\bar{\theta}\right)\right)$ we have

$$vec\left(\bar{\mathbf{\Omega}}_n\left(\bar{\theta}\right)\right) = vec\left(\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\right) + \frac{\gamma_n}{n}\tilde{U}_1 n\left(\bar{\theta}-\hat{\theta}\right) + O_p\left(\frac{\gamma_n}{n^2}\right),$$

where

$$\tilde{U}_1 = \frac{1}{n\gamma_n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\left[l_\tau^{(2)}\left(\hat{\theta}\right)\otimes\mathbf{s}_t\left(\hat{\theta}\right) + \mathbf{s}_\tau\left(\hat{\theta}\right)\otimes l_t^{(2)}\left(\hat{\theta}\right)\right] k\left(\frac{t-\tau}{\gamma_n}\right).$$

Hence, we can get

$$\begin{aligned} P_M &= \mathbf{tr}\left[n\bar{\mathbf{\Omega}}_n\left(\bar{\theta}\right)V\left(\bar{\theta}\right)\right] = nvec\left(\bar{\mathbf{\Omega}}_n\left(\bar{\theta}\right)\right)' vec\left(V\left(\bar{\theta}\right)\right) \\ &= nvec\left(\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\right)' vec\left(V\left(\bar{\theta}\right)\right) + \left[\frac{\gamma_n}{n}\tilde{U}_1 n\left(\bar{\theta}-\hat{\theta}\right)\right]' vec\left(nV\left(\bar{\theta}\right)\right) + O_p\left(\frac{\gamma_n}{n^2}\right). \end{aligned} \tag{119}$$

We can write the second term of (119) as

$$\begin{aligned} &\left[\frac{\gamma_n}{n}\tilde{U}_1 n\left(\bar{\theta}-\hat{\theta}\right)\right]' vec\left(nV\left(\bar{\theta}\right)\right) \\ =\ &\left[\frac{\gamma_n}{n}\tilde{U}_1 n\left(\bar{\theta}-\hat{\theta}\right)\right]'\left[vec\left(-\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right) + O_p\left(\frac{1}{n}\right)\right] \\ =\ &\left[\frac{\gamma_n}{n}\tilde{U}_1\left(-\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}} + \frac{1}{2}\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)' vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right) + O_p\left(\frac{1}{n}\right)\right)\right]' \\ &\times vec\left(-\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right) + O_p(\frac{\gamma_n}{n^2}) \\ =\ &\frac{\gamma_n}{n}vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\tilde{U}_1\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}} \\ &-\frac{1}{2}\frac{\gamma_n}{n}vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\tilde{U}_1\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)' vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right) + O_p\left(\frac{\gamma_n}{n^2}\right), \end{aligned} \tag{120}$$

by (78). And the first term of (119) can be written as

$$\begin{aligned} &nvec\left(\bar{\Omega}_n\left(\hat{\theta}\right)\right)' vec\left(V\left(\bar{\theta}\right)\right) \\ =\ &vec\left(\bar{\Omega}_n\left(\hat{\theta}\right)\right)' nvec\left(V\left(\bar{\theta}\right)\right) \\ =\ &vec\left(\bar{\Omega}_n\left(\hat{\theta}\right)\right)'\left[-vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right) + \frac{1}{n}F_1 + \frac{1}{n}F_2\right] + O_p\left(\frac{1}{n^2}\right) \\ =\ &-\mathbf{tr}\left[\bar{\Omega}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right] + \frac{1}{n}vec\left(\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\right)' F_1 + \frac{1}{n}vec\left(\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\right)' F_2 + O_p\left(\frac{1}{n^2}\right), \end{aligned} \tag{121}$$

by (79). Furthermore, by substituting (80) and (81) into $vec\left(\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\right)' F_1$ and $vec\left(\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\right)' F_2$, we have

$$
\begin{aligned}
& vec\left(\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\right)' F_1 \qquad (122)\\
= \; & -\frac{1}{2}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\otimes vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\hat{\theta}\right)\right]\\
& +\frac{1}{2}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\left(vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\right]\\
& +\frac{1}{2}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right],
\end{aligned}
$$

$$
\begin{aligned}
& vec\left(\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\right)' F_2 \qquad (123)\\
= \; & -\mathbf{tr}\left[\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right)'\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\right]\\
& +\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(2)}}{\hat{p}}\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]-\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\frac{\hat{p}^{(1)}{}'}{\hat{p}}\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]
\end{aligned}
$$

From (121), (122) and (123)

$$
\begin{aligned}
& nvec\left(\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\right)' vec\left(V\left(\overline{\theta}\right)\right)\\
= \; & -\mathbf{tr}\left[\bar{\Omega}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\\
& -\frac{1}{2n}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\otimes vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\hat{\theta}\right)\right]\\
& +\frac{1}{2n}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\left(vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\right]\\
& +\frac{1}{2n}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\\
& -\frac{1}{n}\mathbf{tr}\left[\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right)'\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\right]\\
& +\frac{1}{n}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(2)}}{\hat{p}}\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]-\frac{1}{n}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\frac{\hat{p}^{(1)}{}'}{\hat{p}}\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\\
& +O_p\left(\frac{1}{n^2}\right). \qquad (124)
\end{aligned}
$$

From (120) and (124)

$$
P_M=-\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]+\frac{\gamma_n}{n}C_1^M+\frac{1}{n}C_2^M+O_p\left(\frac{\gamma_n}{n^2}\right), \qquad (125)
$$

where

$$
C_1^M \;=\; vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)'\tilde{U}_1\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}
$$

$$-\frac{1}{2}vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)'\tilde{U}_1\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right),$$

$$\begin{aligned}
C_2^M &= -\frac{1}{2n}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\otimes vec\left(\bar{\mathbf{H}}_n\left(\hat{\theta}\right)^{-1}\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\hat{\theta}\right)\right] \\
&\quad +\frac{1}{2n}\mathbf{tr}\left[\begin{array}{c}\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)' \\ \times\left[\left(vec\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right)'\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\end{array}\right] \\
&\quad +\frac{1}{2n}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)'\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right] \\
&\quad -\frac{1}{n}\mathbf{tr}\left[\left[\left(\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\right)'\otimes\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]\bar{\mathbf{H}}_n^{(3)}\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\right] \\
&\quad +\frac{1}{n}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(2)}}{\hat{p}}\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]-\frac{1}{n}\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\frac{\hat{p}^{(1)}}{\hat{p}}\frac{\hat{p}^{(1)\prime}}{\hat{p}}\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]
\end{aligned}$$

Since in (125) $-\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]=P_T$, we have proved (33) in Theorem 4.1.

From Li et al. (2017),

$$\ln p\left(\mathbf{y}|\overline{\theta}\right)=\ln p\left(\mathbf{y}|\hat{\theta}\right)-\frac{1}{2n}C_{21}+\frac{1}{2n}C_{23}+\frac{1}{8n}C_{12}+O_p\left(n^{-2}\right).$$

Then we have

$$\begin{aligned}
\mathrm{DIC}_M &= -2\ln p\left(\mathbf{y}|\overline{\theta}\right)+2P_M \\
&= -2\ln p\left(\mathbf{y}|\hat{\theta}\right)+\frac{1}{n}C_{21}-\frac{1}{n}C_{23}-\frac{1}{4n}C_{12} \\
&\quad -2\mathbf{tr}\left[\bar{\mathbf{\Omega}}_n\left(\hat{\theta}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\theta}\right)\right]+\frac{2\gamma_n}{n}C_1^M+\frac{2}{n}C_2^M+O_p\left(\frac{\gamma_n}{n^2}\right) \\
&= \mathrm{TIC}+\frac{2\gamma_n}{n}C_1^M+\frac{1}{n}\left(C_{21}-C_{23}-\frac{1}{4}C_{12}+2C_2^M\right)+O_p\left(\frac{\gamma_n}{n^2}\right).
\end{aligned}$$

We have proved (34) in Theorem 4.1. The proof of Theorem 4.1 is completed.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, Springer Verlag, **1**, 267-281.

Andrews, D. W. K. (1987). Consistency in nonlinear econometric models: A generic uniform law of large numbers. *Econometrica*, **55(6)**, 1465-1471.

Andrews, D. W. K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariant Matrix Estimation. *Econometrica*, **59(3)**, 817-858.

Avramov, D., and Zhou, G. F. (2010). Bayesian portfolio analysis. *Annual Review of Financial Economics*, **2**, 25-47.

Bai, J., and Wang, P. (2015). Identification and Bayesian estimation of dynamic factor models. *Journal of Business & Economic Statistics*, **33(2)**, 221-240.

Berg, A., Meyer, R., and Yu, J. (2004). Deviance information criterion for comparing stochastic volatility models. *Journal of Business and Economic Statistics*, **22**, 107-120.

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, **71(356)**, 791-799

Brooks, S. (2002). Discussion on the paper by Spiegelhalter, Best, Carlin, and van de Linde (2002). *Journal of the Royal Statistical Society Series B*, **64**, 616–618.

Burnham, K., and Anderson, D. (2002). *Model Selection and Multi-model Inference: A Practical Information-theoretic Approach*, Springer.

Celeux, G., Forbes, F., Robert, C., and Titterington, D. (2006). Deviance Information Criteria for Missing Data Models. *Bayesian Analysis*, **1**, 651–674.

Chan, J. C., and Grant, A. L. (2016a). On the observed-data deviance information criterion for volatility modeling. *Journal of Financial Econometrics*, **14(4)**, 772-802.

Chan, J. C., and Grant, A. L. (2016b). Fast computation of the deviance information criterion for latent variable models. *Computational Statistics and Data Analysis*, **100**, 847-859.

Chan, J. C., and Eisenstat, E. (2018). Bayesian model comparison for time-varying parameter VARs with stochastic volatility. *Journal of Applied Econometrics*, **33(4)**, 509-532.

Chen, C. (1985). On asymptotic normality of limiting density functions with Bayesian implications. *Journal of the Royal Statistical Society Series B*, **47**, 540–546.

Clark, P. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, **41**, 135-155.

Creal, D. (2012). A survey of sequential Monte Carlo methods for economics and finance. *Econometric reviews*, **31(3)**, 245-296.

Davidson, J. (1992). A central limit theorem for globally nonstationary near-epoch dependent functions of mixing processes. *Econometric theory*, **8**, 313-329.

Davidson, J. (1993). A central limit theorem for globally nonstationary near-epoch dependent functions of mixing processes: The asymptotically degenerate case. *Econometric theory*, **9**, 402-412.

De Jong, R. M. (1997). Central limit theorems for dependent heterogeneous random variables. *Econometric theory*, **13**, 353-367.

De Jong, R. M., and Davidson, J. (2000). Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica*, **68(2)**, 407-423.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, **39**, 1-38.

Ding, J., Tarokh, V., and Yang, Y. (2019) Optimal variable selection in regression models. Working paper, University of Minnesota.

Doucet, A., and Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, **12**, 656-704.

Doucet, A., and Shephard, N. (2012). Robust inference on parameters via particle filters and sandwich covariance matrices. Working Paper, University of Oxford, Department of Economics.

Fama, E. F., and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, **33**, 3-56.

Galán, J. E., Veiga, H., and Wiper, M. P. (2014). Bayesian estimation of inefficiency heterogeneity in stochastic frontier models. *Journal of Productivity Analysis*, **42**, 85-101.

Gallant, A. R., and White, H. (1988). *A unified theory of estimation and inference for nonlinear dynamic models*. Blackwell.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC press.

Geweke, J., and Keane, M. (2001). Computationally intensive methods for integration in econometrics. *Handbook of Econometrics*, **5**, 3463-3568.

Geweke, J., and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, **164**, 130-141.

Ghosh, J., and Ramamoorthi, R. (2003). *Bayesian Nonparametrics*, Springer Verlag.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings F*, **140**, 107–113.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

Ibrahim, J., Zhu, H., and Tang, N. S. (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association*, **103**, 1648–1658.

Inoue, A., and Shintani, M., (2018), Quasi-Bayesian model selection. *Quantitative Economics*, **9**, 1265–97.

Iskrev, N. (2008). Evaluating the information matrix in linearized DSGE models. *Economics Letters*, **99(3)**, 607-610.

Isserlis, L. (1918) On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, **12**, 134–139.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90(430)**, 773-795.

Kass, R., Tierney, L., and Kadane, J. (1990) The validity of posterior expansions based on Laplace's Method. in *Bayesian and Likelihood Methods in Statistics and Econometrics*, ed. by S. Geisser, J.S. Hodges, S.J. Press and A. Zellner. Elsevier Science Publishers B.V.: North-Holland.

Kass, R. E., and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the american statistical association*, **90(431)**, 928-934.

Kim, J. (1994). Bayesian asymptotic theory in a time series model with a possible nonstationary process. *Econometric Theory*, **10**, 764-773.

Kim, J. (1998). Large sample properties of posterior densities, Bayesian information criterion and the likelihood principle in nonstationary time series models. *Econometrica*, **66**, 359-380.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, **65**, 361-393.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, **95(2)**, 391-413.

Li, Y., Yu, J., and Zeng, T. (2017). Deviation information criterion for Bayesian model comparison: justification and variation, Working Paper, Singapore Management University.

Li, Y., Yu, J., and Zeng, T. (2019). Hypothesis Testing, Specification Testing and Model Selection Based on the MCMC Output using R. *Handbook of Statistics*,Vol 41, 81-115.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, **103(481)**, 410-423.

Louis, T. A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, **44**, 226-233.

Magnus, J., and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, Chichester.

Meyer, R., and Yu, J. (2000). BUGS for a Bayesian analysis of stochastic volatility models. *Econometrics Journal*, **3**, 198-215.

Meyn, S. P., and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Millar, R. B. (2009). Comparison of hierarchical Bayesian models for over-dispersed count data using DIC and Bayes factors. *Biometrics*. **65**, 962-969.

Millar, R. B., and McKechnie, S. (2014). A one-step-ahead pseudo DIC for comparison of Bayesian state-space models. *Biometrics*. **70**, 972-980.

Neyman, J., and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*. **16(1)**, 1-32.

Newey, W. K., and West, K. D.(1987). A simple, positive semi-definite, heteroscedasticity and autocorrelation consistent covariance matrix. *Econometrica*. **55**, 703 - 708.

Norets, A. (2009). Inference in dynamic discrete choice models with serially correlated unobserved state variables. *Econometrica*, **77(5)**, 1665-1682.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society B*, **61**, 479-482.

Pitt, M. K., and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, **94**, 590-599.

Poyiadjis, G., Doucet, A., Singh, S. S. (2011). Particle approximations of the score and observed information matrix in state-space models with application to parameter estimation. *Biometrika*, **98(1)**, 65-80.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B*, **71**, 319-392.

Rue, H., Steinsland, I. and Erland, S. (2004). Approximating hidden Gaussian Markov random fields. *Journal of the Royal Statistical Society Series B*, **66**, 877-892.

Schwarz, G. E. (1978), Estimating the dimension of a model. *Annals of Statistics*, **6(2)**, 461–464.

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, **64**, 583–639.

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society Series B*, **76**, 485-493.

Stern, S. (1997). Simulation-based estimation. *Journal of Economic Literature*, **35**, 2006-2039.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39(1)**, 44-47.

Tanner, M., and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528-540.

Takeuchi, K. (1976). Distribution of information statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, **153**, 12-18. [In Japanese]

Vehtari, A., and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, **6**, 142-228.

Vignat, C. (2012). A generalized Isserlis theorem for location mixtures of Gaussian random vectors. *Statistics & Probability Letters*, **82(1)**, 67-71.

White, H. (1996). Estimation, inference and specification analysis. *Cambridge university press.*

Withers, C. S., (1985) The moments of the multivariate normal, *Bulletin of the Australian Mathematical Society*, **32**, 103-107

Wooldridge, J. M. (1994). Estimation and inference for dependent processes. *Handbook of Econometrics*, **4**, 2639-2738.

Yu, J. (2005). On leverage in a stochastic volatility model. *Journal of Econometrics*, **127**, 165-178.

Zhou, G. F. (1993). Asset pricing testing under alternative distributions. *Journal of Finance*, **5**, 1927-1942.